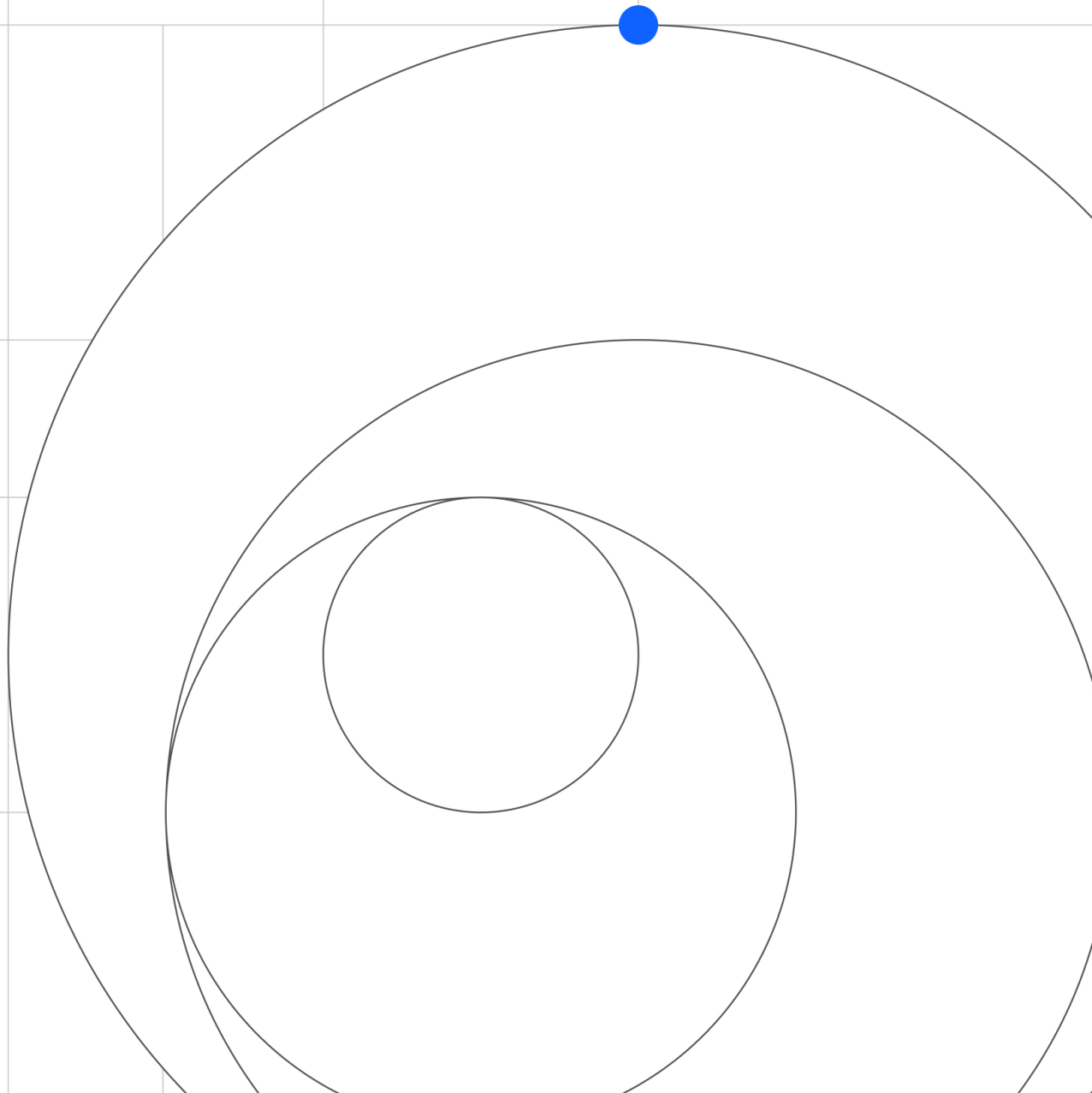


Foundation model: opportunità, rischi e attenuazioni



Riconoscimenti

Con gratitudine verso gli sponsor esecutivi del flusso di lavoro del Comitato etico per l'AI, Christina Montgomery e Francesca Rossi e il contributo dei membri del flusso di lavoro 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 e Suj Perepa.

Sommario

04

Esecutivo
Riepilogo

16

Rischio
Esempi

05

Introduzione

24

Principi, pilastri
e governance

06

Vantaggi dei
Foundation model

25

Guardrail
e attenuazioni

08

Rischi dei
Foundation model

27

Politiche, normative e best
practice in materia di AI
Esempi

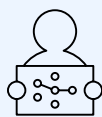
Executive summary

L'ascesa dei foundation model offre alle imprese nuove ed entusiasmanti opportunità, ma solleva anche nuovi e più ampi interrogativi sulla loro progettazione, sviluppo, implementazione e utilizzo etici. Secondo un recente sondaggio dell'IBM Institute for Business Value [sull'AI generativa](#), le organizzazioni esprimono già preoccupazioni su questioni legate alla fiducia, in particolare come ostacoli agli investimenti. Le preoccupazioni principali sono la cybersecurity (57%), la privacy (51%) e l'accuratezza (47%). Molte organizzazioni hanno sul serio queste preoccupazioni prima della *consumerizzazione* dell'AI generativa, esprimendo l'intenzione di investire almeno il 40% in più nell'etica dell'AI nei successivi tre anni. La consapevolezza dei rischi e dei possibili modi per mitigarli è il primo passo fondamentale per costruire sistemi di AI affidabili.

In questo documento:



Esamineremo i vantaggi dei foundation model, tra cui la loro capacità di eseguire compiti impegnativi, la possibilità di accelerare l'adozione dell'AI, la capacità di aumentare la produttività e i vantaggi in termini di costi.



Discuteremo le tre categorie di rischio, compresi i rischi noti derivanti dalle precedenti forme di intelligenza artificiale, i rischi noti amplificati dai modelli di fondazione e i rischi emergenti intrinseci alle capacità generative dei modelli di fondazione.



Illustreremo i principi, i pilastri e la governance che costituiscono la base delle iniziative etiche dell'AI di IBM e indicheremo i limiti per l'attenuazione del rischio.

Introduzione

Poiché l'uso dell'AI continua a espandersi, modelli di intelligenza artificiale ampi e complessi stanno offrendo risultati prestazionali promettenti, oltre a risolvere alcuni dei problemi più impegnativi della società. Tuttavia, la creazione di grandi set di dati di formazione e modelli complessi per ogni applicazione di intelligenza artificiale può comportare oneri significativi per le aziende. I modelli di fondazione forniscono un percorso per ottenere il meglio da entrambi i mondi: essi creano modelli potenti e all'avanguardia, riutilizzandoli direttamente o applicando metodi di ottimizzazione per implementare una varietà di casi d'uso, anziché addestrare nuovi modelli per ogni caso d'uso. Ad esempio, IBM®Research ha sviluppato [modelli di fondazione per l'ispezione visiva](#). Questi modelli di fondazione apprendono la rappresentazione generale delle superfici e delle piste in calcestruzzo e possono essere ulteriormente ottimizzati per casi d'uso specifici come il rilevamento di crepe o l'ispezione dei difetti con una minor quantità di dati etichettati.

IBM definisce un *modello di fondazione* nella forma di un modello di intelligenza artificiale che può essere adattato a un'ampia gamma di attività downstream. I modelli di fondazione sono in genere modelli generativi su larga scala che vengono addestrati su dati non etichettati utilizzando la supervisione automatica. Trattandosi di modelli su larga scala, i modelli di fondazione possono includere miliardi di parametri.

IBM è un'azienda che offre cloud ibrido e AI e che gode di una lunga reputazione come gestore responsabile dei dati impegnato nell'[etica dell'AI](#). Sfruttando la forza dei nostri team di [ricerca](#), [prodotto](#) e [consulenza](#), insieme a partner esterni come [Hugging Face](#), contribuiamo a portare la potenza dei modelli di fondazione ai nostri clienti e a sviluppare un'AI affidabile in ogni azienda. IBM continua inoltre a investire nella creazione di nuove piattaforme, come la piattaforma e le tecnologie di dati e AI di [IBM® watsonx™](#), per la progettazione e lo sviluppo di modelli di AI che si comportino in maniera verificabile e affidabile.

Questo documento descrive il punto di vista di IBM sull'etica dei modelli di fondazione. Si tratta della prima versione di questo documento. Le successive espanderanno vari aspetti dell'approccio etico del modello di fondazione di IBM. Ci auguriamo che questo documento sia utile per tutti gli stakeholder nello sviluppo, nell'implementazione e nell'utilizzo responsabile del modello di fondazione.

Vantaggi dei Foundation model

I modelli di fondazione possono migliorare in maniera significativa il processo di sviluppo dei sistemi di intelligenza artificiale e contribuire a far progredire l'AI dalla fase di esplorazione a quella di adozione nelle imprese. I vantaggi includono:

Esecuzione di compiti complessi

I modelli di fondazione mostrano un aumento significativo delle prestazioni nella risoluzione di problemi difficili e complessi. Ad esempio, il [modello di fondazione geospaziale](#) della [collaborazione tra IBM e NASA](#) è progettato per convertire i dati satellitari della NASA in mappe di disastri naturali come inondazioni e altri cambiamenti ambientali. Il modello potrebbe anche essere usato per: aiutare a rivelare il passato del nostro pianeta; stimare i rischi per le colture, le aziende o le infrastrutture a causa di condizioni meteorologiche avverse; sviluppare strategie di adattamento ai cambiamenti climatici; assistere il settore agroalimentare. Il modello sarà reso disponibile in anteprima ai clienti IBM attraverso [IBM Environmental Intelligence Suite](#).

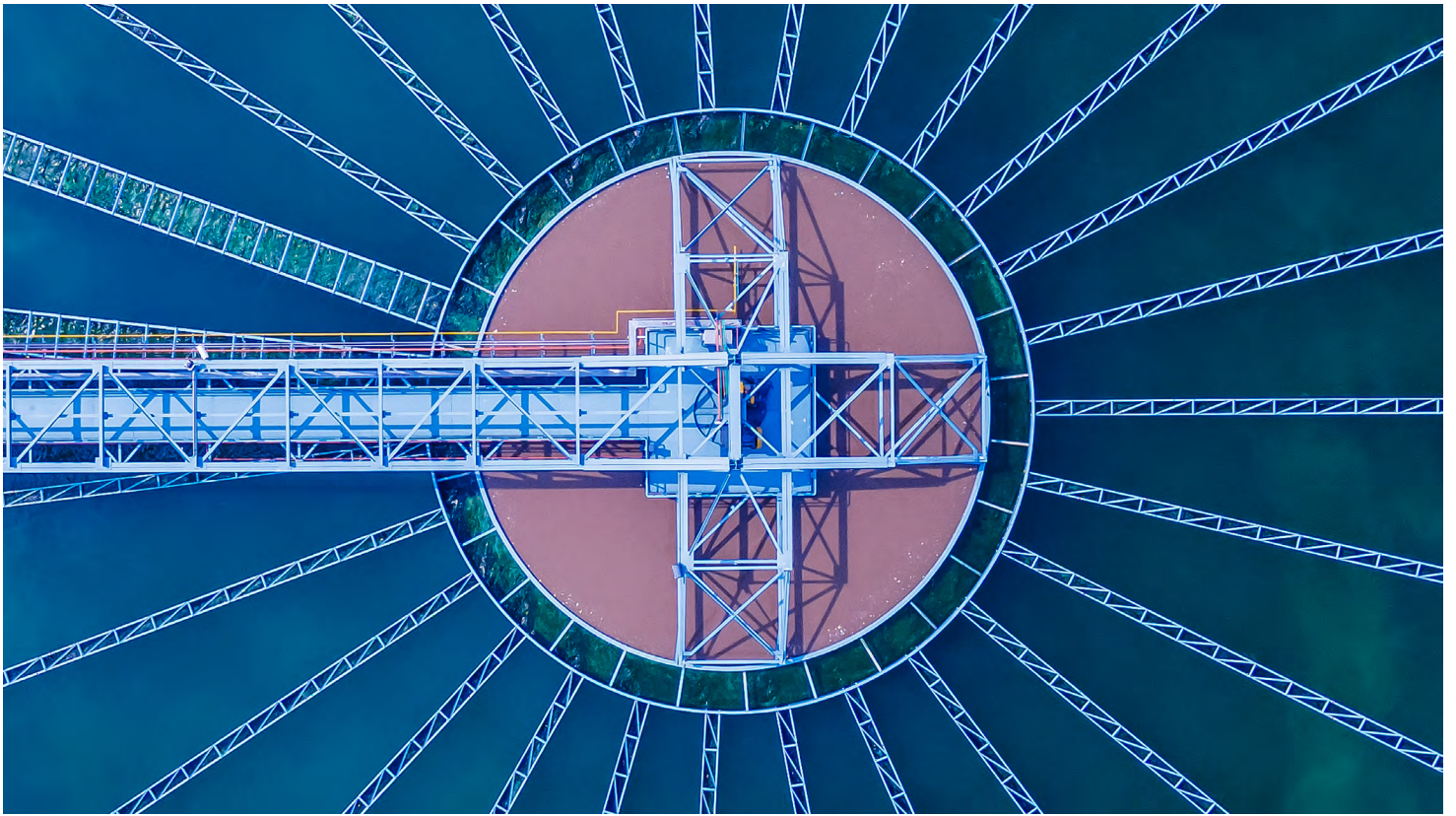
Come altro esempio, [MoLFormer-XL](#) di IBM è un modello di fondazione che deduce la struttura delle molecole a partire da semplici rappresentazioni e facilita l'apprendimento di varie attività downstream come la previsione delle proprietà fisiche e quantistiche di una molecola, l'identificazione di molecole simili, lo screening di molecole già approvate per nuovi casi d'uso e la scoperta di nuove molecole. [Moderna e IBM](#) stanno studiando i modi possibili in cui utilizzare MoLFormer per prevedere le proprietà delle molecole e comprendere le caratteristiche di potenziali farmaci a mRNA.

Maggiore produttività

La natura generativa dei modelli di fondazione espande il numero di aree in cui l'AI può essere utilizzata in azienda con lo scopo di migliorare la produttività attraverso l'automazione delle tediose attività di routine, consentendo agli utenti di dedicare più tempo alle attività creative e innovative. Ad esempio, [IBM Watsonx® Code Assistant](#), supportato da [modelli di fondazione](#), consente agli sviluppatori di ogni livello di esperienza di scrivere codici utilizzando consigli generati dall'AI.

Time-to-value più rapido

I modelli di fondazione vengono generalmente addestrati con dati non etichettati, più accessibili in quantità maggiori rispetto ai dati etichettati. Una volta addestrati, i modelli di fondazione possono essere utilizzati direttamente o dopo essere stati ottimizzati per le applicazioni downstream, utilizzando una piccola quantità di dati etichettati specializzati in grado di ridurre il time-to-value del processo di creazione.



Utilizzare diverse modalità di dati

I modelli di fondazione possono essere addestrati utilizzando varie modalità di dati, come linguaggio naturale, testo, immagini e audio. Possono anche essere applicati ad attività che richiedono diversi tipi di dati, come dati della serie del tempo, dati geospaziali, dati in formato tabellare, dati semistrutturati e dati a modalità mista come il testo combinato con immagini.

Spese ammortizzate

Sebbene il costo iniziale della formazione di un modello di fondazione sia significativamente superiore rispetto al training di un modello di intelligenza artificiale tradizionale, il costo incrementale della sua applicazione a una nuova attività è notevolmente inferiore. L'utilizzo di modelli di fondazione pre-addestrati potrebbe aiutare a eliminare la necessità che le imprese effettuino investimenti sostanziali per addestrare i modelli di fondazione a sperimentare le loro nuove capacità. Per un'azienda, l'affidabilità dei modelli, l'efficienza energetica, le prestazioni, la portabilità e la capacità di utilizzare i dati aziendali in modo efficace e sicuro sono fondamentali.

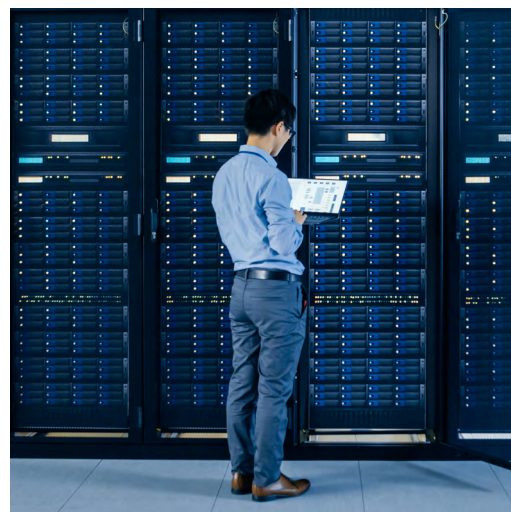
IBM consente alle imprese di creare e possedere il valore dei modelli di fondazione per il proprio business apportando le migliori innovazioni dalla comunità globale e aperta dell'AI, consentendo l'esecuzione efficiente in ambienti di elaborazione ibridi, aiutando a mitigare i rischi e governando rigorosamente l'AI.

Rischi dei Foundation model

Come tutte le tecnologie in rapida evoluzione, i foundation model presentano anche dei rischi oltre ai benefici. Alcuni sono rischi di tipo legale, ad esempio restrizioni sul trasferimento o l'utilizzo dei dati, e devono essere attentamente valutati in base alla legislazione attuale e in evoluzione. Altri rischi hanno una natura etica e devono essere considerati attentamente in modo che la tecnologia abbia un impatto positivo. In generale, i rischi dell'AI sollevano dubbi socio-tecnici e devono essere dunque affrontati e mitigati attraverso metodi socio-tecnici, tra cui strumenti software, processi di valutazione del rischio, framework di etica dell'AI, meccanismi di governance, consultazioni multistakeholder, standard e normative. Elencheremo i rischi considerando le seguenti 3 categorie:

1. **Tradizionale:** rischi noti derivanti da forme precedenti dei sistemi di intelligenza artificiale.
2. **Amplificato:** rischi noti ma ora intensificati a causa di caratteristiche intrinseche dei modelli di fondazione, in particolare, le loro capacità generative intrinseche.
3. **Nuovo:** rischi emergenti intrinseci ai foundation model le loro capacità generative intrinseche.

Inoltre, strutturiamo l'elenco dei rischi a seconda che siano principalmente associati al contenuto fornito al foundation model (l'input) o al contenuto generato da questo (l'output) oppure che siano correlati a difficoltà aggiuntive.



1. Rischi associati all'input

Fase di formazione e ottimizzazione

Gruppo	Rischio	Perché questo rappresenta un problema?	Indicatore
Equità	Distorsione dei dati: distorsioni storiche, rappresentative e sociali presenti nei dati utilizzati per addestrare e mettere a punto il modello.	Addestrare un sistema AI su dati con distorsioni, come ad esempio distorsioni storiche o rappresentative, potrebbe portare a output distorti o alterati che possono rappresentare ingiustamente o anche discriminare determinati gruppi o individui. Oltre agli impatti sociali negativi, le entità aziendali potrebbero anche incorrere in conseguenze legali, interruzioni nelle operazioni o danni alla reputazione derivanti da risultati distorti dei modelli.	Amplificato
Affidabilità	Avvelenamento dei dati: un tipo di attacco in cui un avversario o un insider malevolo inserisce campioni intenzionalmente corrotti, falsi, fuorvianti o errati nel set di dati di addestramento o di perfezionamento.	L'avvelenamento dei dati può rendere il modello sensibile a un modello di dati dannoso e produrre l'output desiderato dall'avversario. Può creare un rischio per la sicurezza in cui gli avversari possono forzare il comportamento del modello a proprio vantaggio. Oltre a produrre risultati involontari e potenzialmente dannosi, un disallineamento del modello derivante dall'avvelenamento dei dati può comportare per le entità aziendali conseguenze legali, interruzioni nelle operazioni o danni alla reputazione.	Tradizionale
Allineamento del valore	Data curation: quando i dati di addestramento o di ottimizzazione vengono raccolti o preparati in modo improprio.	Una data curation non corretta può influire negativamente sul modo in cui un modello viene addestrato, con conseguente comportamento non conforme rispetto ai valori previsti. Esempi di data curation non corretta potrebbero includere errori di etichettatura o annotazione nei dati utilizzati per l'addestramento o la messa a punto del modello. Correggere i problemi dopo che il modello è stato addestrato e implementato potrebbe non essere sufficiente a garantire un comportamento corretto. Un modello di comportamento non corretto può comportare per le entità aziendali conseguenze legali, interruzioni nelle operazioni o danni alla reputazione.	Amplificato
	Riqualificazione basata a valle: utilizzo di output indesiderati (imprecisi, inappropriati, contenuti dell'utente, ecc.) dalle applicazioni a valle a scopi di nuovi addestramenti.	Il riutilizzo dell'output a valle per il nuovo addestramento di un modello senza l'implementazione di un adeguato controllo dell'uomo aumenta le possibilità che gli output indesiderati vengano incorporati nei dati di addestramento o di messa a punto del modello, generando potenzialmente un output ancora più indesiderato. Un modello di comportamento non corretto può provocare conseguenze legali o danni alla reputazione per le entità aziendali. Il mancato rispetto delle leggi sul trasferimento dei dati potrebbe comportare multe e altre conseguenze legali.	Nuovo
Leggi sui dati	Trasferimento dei dati: la legge e altre restrizioni possono limitare o vietare il trasferimento dei dati.	Le restrizioni al trasferimento dei dati possono influire sulla disponibilità dei dati necessari per l'addestramento di un modello AI e possono portare a dati scarsamente rappresentati. Oltre all'impatto sulla disponibilità dei dati, il mancato rispetto delle leggi e dei regolamenti sul trasferimento dei dati potrebbe comportare multe e altre conseguenze legali.	Tradizionale
	Utilizzo dei dati: la legge e altre restrizioni possono limitare o vietare l'utilizzo di alcuni dati per specifici casi d'uso di AI.	Il mancato rispetto delle leggi e dei regolamenti sull'utilizzo dei dati potrebbe comportare multe e altre conseguenze legali.	Tradizionale
	Acquisizione di dati: leggi e altri regolamenti potrebbero limitare la raccolta di determinati tipi di dati per specifici casi d'uso di AI.	Il mancato rispetto delle leggi e dei regolamenti sull'acquisizione di dati potrebbe comportare multe e altre conseguenze legali.	Amplificato

Gruppo	Rischio	Perché questo rappresenta un problema?	Indicatore
Proprietà intellettuale	Diritti di utilizzo dei dati: i termini di servizio, le leggi sul copyright, la conformità alle licenze o altri problemi relativi alla proprietà intellettuale possono limitare la possibilità di utilizzare determinati dati per la creazione di modelli.	Le leggi e i regolamenti riguardanti l'uso dei dati per addestrare l'AI sono instabili e possono variare da paese a paese, il che crea problemi nello sviluppo di modelli. Se l'utilizzo dei dati viola regole o restrizioni, le entità aziendali potrebbero incorrere in multe, danni alla reputazione, interruzioni nelle operazioni e altre conseguenze legali.	Amplificato
Trasparenza	Trasparenza dei dati: sfida nel documentare il modo in cui i dati di un modello sono stati raccolti, resi accurati e utilizzati per addestrare un modello.	La trasparenza dei dati è importante per la conformità legale e l'etica dell'AI. Le informazioni mancanti limitano la capacità di valutare i rischi associati ai dati. La mancanza di requisiti standardizzati potrebbe limitare la divulgazione, in quanto le organizzazioni proteggono i segreti commerciali e cercano di impedire ad altri di copiare i propri modelli.	Amplificato
	Provenienza dei dati: sfida relativa alla standardizzazione e alla definizione di metodi per verificare la provenienza dei dati.	Non tutte le fonti di dati sono affidabili. I dati potrebbero essere stati raccolti, manipolati o falsificati in modo non etico. L'utilizzo di dati inaffidabili può provocare comportamenti indesiderati nel modello. Le entità aziendali potrebbero incorrere in multe, danni alla reputazione, interruzioni nelle operazioni e altre conseguenze legali.	Amplificato
Privacy	Informazioni personali nei dati: inclusione o presenza di informazioni di identificazione personale (PII) e informazioni personali sensibili (SPI) nei dati utilizzati per l'addestramento o l'ottimizzazione del modello.	Se non viene sviluppato correttamente per proteggere i dati sensibili, il modello potrebbe esporre informazioni personali nell'output generato. Inoltre, i dati personali o sensibili devono essere esaminati e gestiti in conformità con le leggi e i regolamenti sulla privacy. In caso di violazione, le entità aziendali possono incorrere in multe, danni alla reputazione, interruzioni nelle operazioni e altre conseguenze legali.	Tradizionale
	Reidentificazione: anche con l'eliminazione delle informazioni di identificazione personale (PII) e delle informazioni personali sensibili (SPI) dai dati, potrebbe essere comunque possibile identificare le persone grazie ad altre funzioni disponibili nei dati.	I dati che possono rivelare informazioni personali o sensibili devono essere esaminati nel rispetto delle leggi e dei regolamenti sulla privacy, in quanto le entità aziendali potrebbero incorrere in multe, danni alla reputazione, interruzioni nelle operazioni e altre conseguenze legali se riscontrate in violazione.	Tradizionale
	Diritti sulla privacy dei dati: sfide relative alla capacità di fornire diritti agli interessati come la rinuncia, il diritto di accesso, il diritto all'oblio.	L'identificazione o l'uso improprio dei dati potrebbe comportare la violazione delle leggi sulla privacy. Un utilizzo improprio o una richiesta di rimozione dei dati potrebbe costringere le organizzazioni ad addestrare nuovamente il modello, il che è costoso. Inoltre, le entità aziendali potrebbero incorrere in multe, danni alla reputazione, interruzioni nelle operazioni e altre conseguenze legali in caso di mancato rispetto delle norme e dei regolamenti sulla privacy dei dati.	Amplificato
	Consenso informato: dati raccolti per l'addestramento dei modelli AI senza il consenso informato del titolare anche quando ciò è consentito dalla legge.	In determinate circostanze, potrebbe non essere etico raccogliere e utilizzare i dati senza il consenso della persona. Questo utilizzo comporta anche possibili rischi per la reputazione.	Tradizionale

Inferenza Fase

Gruppo	Rischio	Perché questo rappresenta un problema?	Indicatore
Privacy	Informazioni personali nel prompt: divulgazione di informazioni personali o informazioni personali sensibili come parte del prompt inviato al modello.	I dati richiesti potrebbero essere memorizzati o utilizzati successivamente per altri scopi come la valutazione e il nuovo addestramento del modello. Questi tipi di dati devono essere esaminati nel rispetto delle leggi e dei regolamenti sulla privacy. Senza un utilizzo e un data storage adeguati, le entità aziendali potrebbero incorrere in multe, danni alla reputazione, interruzioni nelle operazioni e altre conseguenze legali.	Nuovo
Proprietà intellettuale	Informazioni IP nel prompt: divulgazione di informazioni sul copyright o altre informazioni IP come parte del prompt inviato al modello.	I dati richiesti potrebbero essere memorizzati o utilizzati successivamente per altri scopi come la valutazione e il nuovo addestramento del modello. Questi tipi di dati devono essere esaminati nel rispetto delle leggi e dei regolamenti sulla proprietà intellettuale. Senza un utilizzo e un data storage adeguati, le entità aziendali potrebbero incorrere in multe, danni alla reputazione, interruzioni nelle operazioni e altre conseguenze legali.	Nuovo
	Dati riservati nel prompt: inclusione di dati riservati come parte del prompt inviato al modello.	Se non viene sviluppato correttamente per proteggere i dati riservati, il modello potrebbe esporre informazioni riservate o IP nell'output generato. Inoltre, le informazioni riservate degli utenti finali potrebbero essere raccolte e memorizzate involontariamente.	Nuovo
Affidabilità	Attacco di evasione: tentativo di far sì che un modello produca output non corretti alterando i dati inviati al modello addestrato.	Gli attacchi di evasione alterano il comportamento del modello, solitamente a vantaggio dell'aggressore. Nel caso in cui non vengano tenuti correttamente in considerazione i risultati di output, le entità aziendali potrebbero incorrere in multe, danni alla reputazione, interruzioni nelle operazioni e altre conseguenze legali.	Amplificato
	Attacchi basati su prompt: attacchi avversari come prompt injection (tentativo di forzare un modello a produrre output inaspettato), prompt leaking (tentativi di estrarre il prompt di sistema di un modello), jailbreak (tentativi di sfondare i guardrail stabiliti nel modello) e priming del prompt (tentativo di forzare un modello a produrre un output allineato al prompt).	A seconda del contenuto rivelato, le entità aziendali potrebbero incorrere in multe, danni alla reputazione, interruzioni nelle operazioni e altre conseguenze legali.	Nuovo

2. Rischi associati all'output

Gruppo	Rischio	Perché questo rappresenta un problema?	Indicatore
Equità	Distorsione dell'output: i contenuti generati potrebbero rappresentare ingiustamente determinati gruppi o individui.	Le distorsioni possono danneggiare gli utenti dei modelli AI e amplificare i comportamenti discriminatori esistenti. Le entità aziendali possono incorrere in danni alla reputazione, interruzioni nelle operazioni e altre conseguenze.	Nuovo
	Distorsione decisionale: quando un gruppo è ingiustamente avvantaggiato rispetto a un altro a causa delle decisioni prese da esseri umani utilizzando l'output del modello.	Le distorsioni possono danneggiare le persone interessate dalle decisioni del modello. Le entità aziendali potrebbero incorrere in multe, danni alla reputazione, interruzioni nelle operazioni e altre conseguenze legali.	Tradizionale
Proprietà intellettuale	Violazione del copyright: quando un modello genera contenuti troppo simili o identici a un'opera esistente protetta da copyright o da un contratto di licenza open source.	Le leggi e i regolamenti riguardanti l'uso di contenuti che sembrano uguali o molto simili ad altri dati protetti da copyright sono in gran parte instabili e possono variare da paese a paese, creando problemi nella determinazione e nell'implementazione della conformità. Le entità aziendali potrebbero incorrere in multe, danni alla reputazione, interruzioni nelle operazioni e altre conseguenze legali.	Nuovo
Allineamento del valore	Allucinazione: generazione di contenuti fattuali inesatti o non veritieri.	Gli output falsi possono fuorviare gli utenti ed essere incorporati in artefatti a valle, diffondendo ulteriormente la disinformazione. Ciò può danneggiare sia i proprietari che gli utenti dei modelli AI. Le entità aziendali potrebbero inoltre incorrere in multe, danni alla reputazione, interruzioni nelle operazioni e altre conseguenze legali.	Nuovo
	Output dannoso: quando il modello produce contenuti che incitano all'odio, offensivi e profani (HAP) oppure osceni.	I contenuti che incitano all'odio, offensivi e profani (HAP) oppure osceni possono avere un impatto negativo e danneggiare le persone che interagiscono con il modello. Le entità aziendali potrebbero inoltre incorrere in multe, danni alla reputazione, interruzioni nelle operazioni e altre conseguenze legali.	Nuovo
	Consigli pericolosi: quando un modello fornisce consigli senza avere informazioni sufficienti, con il risultato di un possibile pericolo se il consiglio viene seguito.	Una persona potrebbe agire sulla base di consigli incompleti o preoccuparsi di una situazione che non è applicabile a causa della natura eccessivamente generalizzata del contenuto generato.	Nuovo
Uso improprio	Diffusione di disinformazioni: utilizzo di un modello per creare informazioni fuorvianti o false per ingannare o influenzare un destinatario specifico.	La diffusione di disinformazioni potrebbe influire sulla capacità di un essere umano di prendere decisioni informate. Le entità aziendali potrebbero incorrere in multe, danni alla reputazione, interruzioni nelle operazioni e altre conseguenze legali.	Nuovo
	Contenuti dannosi: utilizzo di un modello per generare contenuti che incitano all'odio, offensivi e profani (HAP) oppure osceni.	I contenuti dannosi potrebbero influire negativamente sul benessere dei suoi destinatari. Le entità aziendali potrebbero incorrere in multe, danni alla reputazione, interruzioni nelle operazioni e altre conseguenze legali.	Nuovo
	Uso non consensuale: utilizzo di un modello per imitare le persone attraverso video (deepfake), immagini, audio o altre modalità senza il loro consenso.	I deepfake possono diffondere disinformazioni su una persona, con conseguente impatto negativo sulla reputazione della persona stessa. Le entità aziendali potrebbero incorrere in multe, danni alla reputazione, interruzioni nelle operazioni e altre conseguenze legali.	Amplificato

Gruppo	Rischio	Perché questo rappresenta un problema?	Indicatore
	Utilizzo pericoloso: utilizzo di un modello con l'unica intenzione di danneggiare le persone.	Le entità aziendali potrebbero incorrere in multe, danni alla reputazione, interruzioni nelle operazioni e altre conseguenze legali.	Nuovo
	Mancata divulgazione: la mancata divulgazione di un contenuto generato da un modello AI.	La mancata divulgazione dei contenuti creati dall'AI può essere considerata ingannevole con conseguente diminuzione della fiducia. L'inganno intenzionale potrebbe comportare una diminuzione dell'azione umana, multe, danni alla reputazione e altre conseguenze legali.	Nuovo
	Utilizzo improprio: utilizzo di un modello per uno scopo per il quale il modello non è stato progettato.	Riutilizzare un modello senza conoscerne i dati originali, l'intento progettuale e gli obiettivi potrebbe causare comportamenti del modello imprevisi e indesiderati.	Amplificato
Generazione di codice dannoso	Generazione di codice dannoso: i modelli possono generare un codice che, quando eseguito, causa danni o influisce involontariamente su altri sistemi.	L'esecuzione di un codice dannoso potrebbe causare vulnerabilità nei sistemi IT. Le entità aziendali potrebbero incorrere in multe, danni alla reputazione, interruzioni nelle operazioni e altre conseguenze legali.	Nuovo
Fiducia mal riposta	Eccessivo/scarso affidamento: quando una persona ripone troppa fiducia o troppa poca fiducia nelle indicazioni di un modello AI.	Nei compiti in cui gli esseri umani effettuano scelte indirizzate dai suggerimenti basati sull'AI, l'affidamento eccessivo o scarso può portare a un processo decisionale inadeguato a causa della fiducia malriposta nel sistema AI, con conseguenze negative che aumentano con l'importanza della decisione. Decisioni sbagliate possono danneggiare le persone e causare danni finanziari, alla reputazione, interruzioni nelle operazioni e altre conseguenze legali per le entità aziendali.	Amplificato
Privacy	Esposizione di informazioni personali: quando le informazioni di identificazione personale (PII) o le informazioni personali sensibili (SPI) vengono utilizzate nei dati di addestramento, di ottimizzazione o come parte del prompt, i modelli potrebbero rivelare tali dati nell'output generato.	La condivisione della proprietà intellettuale delle persone ha un impatto sui loro diritti e li rende più vulnerabili. Inoltre, i dati di output devono essere esaminati rispetto alle leggi e ai regolamenti sulla privacy, poiché le entità aziendali potrebbero incorrere in multe, danni alla reputazione, interruzioni nelle operazioni e altre conseguenze legali se riscontrate in violazione della privacy dei dati o delle leggi sull'utilizzo.	Nuovo
Attendibilità	Output inspiegabile: difficoltà nello spiegare il motivo per cui è stato generato l'output del modello.	I foundation model si basano su complesse architetture di deep learning, rendendo difficile la spiegazione degli output. Senza spiegazioni chiare sull'output del modello, è difficile per gli utenti, i validatori del modello e i revisori comprendere e fidarsi del modello. La mancanza di trasparenza potrebbe avere conseguenze legali in settori altamente regolamentati. Spiegazioni sbagliate potrebbero portare a un'eccessiva fiducia.	Amplificato
Tracciabilità	Attribuzione inaffidabile delle fonti: difficoltà nel determinare da quali dati di addestramento o perfezionamento il modello ha generato una parte o tutto il suo output.	L'incapacità di tracciare l'origine o la provenienza dell'output rende difficile per gli utenti, i validatori del modello e i revisori comprendere e fidarsi del modello.	Nuovo

3. Difficoltà

Gruppo	Rischio	Perché questo rappresenta un problema?	Indicatore
Governance	Trasparenza del modello: la mancanza di trasparenza del modello o l'insufficiente documentazione del processo di sviluppo del modello rende difficile capire come e perché un modello è stato costruito e chi lo ha costruito, aumentando così la possibilità di un uso improprio non intenzionale del modello.	La trasparenza è importante per la conformità legale, l'etica dell'AI e per promuovere l'uso appropriato dei modelli. La mancanza di informazioni potrebbe rendere più difficile la valutazione dei rischi, la modifica del modello o il riutilizzo. Anche la conoscenza di chi ha creato un modello può essere un fattore importante per decidere se fidarsi o meno.	Tradizionale
	Responsabilità: il processo di sviluppo di foundation model è complesso e contiene molti dati, processi e ruoli. Quando l'output del modello non funziona come previsto, può essere difficile determinare la causa principale e assegnare la responsabilità.	Senza documentare adeguatamente le decisioni e assegnare la responsabilità, potrebbe non essere possibile determinare la responsabilità per comportamenti imprevisti o uso improprio.	Amplificato
Conformità legale	Responsabilità legale: determinare chi è responsabile del foundation model.	Se la proprietà o la responsabilità per lo sviluppo del modello è incerta, le autorità di regolamentazione e altri soggetti potrebbero manifestare preoccupazioni sul modello in quanto non sarà chiaro chi è (o chi dovrebbe essere) responsabile dei problemi che lo riguardano o chi può rispondere a domande su di esso. Gli utenti di modelli senza una proprietà definita chiaramente potrebbero incontrare difficoltà nel rispetto della futura regolamentazione AI.	Nuovo
	Proprietà dei contenuti generati: determinazione della proprietà dei contenuti generati dall'AI.	Le leggi e i regolamenti relativi alla proprietà dei contenuti generati dall'AI sono ancora poco definiti e possono variare da paese a paese. Le entità aziendali potrebbero incorrere in multe, danni alla reputazione, interruzioni nelle operazioni e altre conseguenze legali.	Nuovo
	IP dei contenuti generati: incertezza giuridica sui diritti di proprietà intellettuale relativi ai contenuti generati.	Le leggi e i regolamenti sulla determinazione della tutela del diritto d'autore e della brevettabilità dei contenuti generati dall'AI sono ancora poco definiti e possono variare da paese a paese. Le entità aziendali potrebbero incorrere in multe, danni alla reputazione, interruzioni nelle operazioni e altre conseguenze legali se il contenuto generato è protetto da diritti di proprietà intellettuale.	Nuovo
	Attribuzione della fonte: determinare la provenienza del contenuto generato.	Se il modello genera un output identico ai dati utilizzati per addestrare il modello, dovrebbe fornire la provenienza di tale output. In caso contrario, le entità aziendali che implementano o utilizzano il modello possono essere esposte a rischi legali.	Amplificato
Sociale Impatto	Impatto sull'occupazione: l'adozione diffusa di sistemi di AI basati su foundation model potrebbe portare alla perdita di posti di lavoro delle persone, in quanto il loro lavoro è automatizzato, se non vengono riqualificate.	La perdita del lavoro potrebbe portare a una perdita di reddito e quindi un impatto negativo sulla società e sul benessere della persona. La riqualificazione potrebbe essere un problema dato il ritmo dell'evoluzione della tecnologia.	Amplificato

Gruppo	Rischio	Perché questo rappresenta un problema?	Indicatore
	Sfruttamento umano: utilizzo del lavoro fantasma nei modelli AI di addestramento, condizioni di lavoro inadeguate, mancanza di assistenza sanitaria anche mentale, compensi iniqui.	Il foundation model dipende ancora dal lavoro umano per reperire, gestire e progettare i dati utilizzati per addestrare il modello. Lo sfruttamento umano per queste attività potrebbe avere un impatto negativo sulla società e sul benessere della persona. Inoltre, le entità aziendali potrebbero incorrere in multe, rischi alla reputazione, interruzioni nelle operazioni e altre conseguenze legali.	Amplificato
	Impatto sull'ambiente: aumento delle emissioni di carbonio e dell'utilizzo di acqua per addestrare e gestire modelli di AI.	Il consumo di grandi quantità di energia per l'addestramento dell'AI contribuisce alle emissioni di carbonio che potrebbero accelerare il cambiamento climatico. Le risorse idriche utilizzate per il raffreddamento dei server dei data center dell'AI non possono più essere destinate ad altri usi necessari.	Amplificato
	Impatto sulla diversità culturale: i sistemi AI potrebbero rappresentare eccessivamente determinate culture con una conseguente omogeneizzazione della cultura e dei pensieri.	Le lingue, i punti di vista e le istituzioni dei gruppi sottorappresentati potrebbero essere soppressi, riducendo così la diversità di pensiero e di cultura.	Nuovo
	Impatto sulle azioni umane: disinformazione e informazioni non corrette generate dai foundation model, inclusa la generazione di contenuti manipolativi.	L'AI può generare disinformazione che sembra reale. Pertanto, le persone potrebbero non riconoscerla come falsa informazione. Inoltre, può semplificare la capacità di attori nefasti di generare contenuti con l'intenzione di manipolare i pensieri e il comportamento umano.	Amplificato
	Impatto sulla formazione, elusione dell'apprendimento: utilizzo di modelli AI per eludere il processo di apprendimento.	I modelli AI facilitano la ricerca rapida di soluzioni o la risoluzione di problemi complessi. Questi sistemi possono essere utilizzati in modo improprio dagli studenti per eludere il processo di apprendimento. La facilità di accesso a questi modelli fa sì che gli studenti abbiano una comprensione superficiale dei concetti e ostacola una formazione più approfondita che potrebbe basarsi sulla comprensione di tali concetti.	Nuovo
	Impatto sulla formazione. Plagio: utilizzo di modelli AI per plagiare in maniera intenzionale o involontaria lavori esistenti.	I modelli AI possono essere utilizzati per rivendicare la paternità o l'originalità di opere create da altre persone, commettendo così un plagio. Rivendicare il lavoro altrui come proprio è immorale e spesso illegale.	Nuovo

Esempi di rischio

Forniamo esempi riportati dalla stampa per aiutare a spiegare molti dei rischi dei foundation model. Molti di questi eventi trattati dalla stampa sono ancora in evoluzione o sono stati risolti, e fare riferimento a essi può aiutare il lettore a comprendere i potenziali rischi e lavorare per mitigarli. Mettere in evidenza questi esempi è solo a scopo illustrativo.

Esempi di rischio: input

Addestramento e messa a punto Fase

Gruppo	Rischio	Esempio
Equità	Distorsione dei dati: distorsioni storiche, rappresentative e sociali presenti nei dati utilizzati per addestrare e mettere a punto il modello.	Distorsioni nel settore sanitario La ricerca sul rafforzamento delle disuguaglianze in medicina mette in evidenza che l'utilizzo dei dati e dell'AI per trasformare il modo in cui le persone ricevono assistenza sanitaria è tanto effettivo quanto i dati alla base di essi, il che significa che l'uso di dati di addestramento con scarsa rappresentanza di minoranze o che riflettono ciò che è già diseguale nell'assistenza sanitaria può portare a un aumento delle disuguaglianze nel settore sanitario. [Forbes, dicembre 2022]
Allineamento del valore	Riqualificazione basata a valle: utilizzo di output indesiderabili (imprecisi, inappropriati, contenuti dell'utente, ecc.) dall'applicazione a valle per scopi di riqualificazione	Collasso del modello dovuto all'addestramento tramite contenuti generati dall'AI Come affermato nell'articolo originale, un gruppo di ricercatori ha studiato il problema dell'utilizzo di contenuti generati dall'AI per l'addestramento invece di contenuti generati dall'uomo. Questo gruppo ha scoperto che i grandi modelli linguistici alla base della tecnologia potrebbero potenzialmente essere addestrati su altri contenuti generati dall'AI mentre continua a diffondersi in massa su Internet, un fenomeno che hanno coniato come "collasso del modello". [Business Insider, agosto 2023]
Leggi sui dati	Trasferimento dei dati: la legge e altre restrizioni possono limitare o vietare il trasferimento dei dati.	Leggi sulla restrizione dei dati Come affermato nell'articolo della ricerca, le misure di localizzazione dei dati che limitano la capacità di spostare i dati a livello globale ridurranno la potenzialità di sviluppare capacità AI su misura. Questo influirà direttamente sull'AI fornendo meno dati di addestramento e minando indirettamente gli elementi costitutivi su cui è sviluppata l'AI. Gli esempi includono le restrizioni GDPR sul trattamento e l'uso dei dati personali. [Brookings, dicembre 2018]
Proprietà intellettuale	Diritti di utilizzo dei dati: i termini di servizio, le leggi sul copyright, la conformità delle licenze o altri problemi relativi alla proprietà intellettuale possono limitare la possibilità di utilizzare determinati dati per la creazione di modelli.	Reclami per violazione del copyright del testo Secondo l'articolo originale, il New York Times ha citato in giudizio OpenAI e Microsoft accusandoli di utilizzare milioni di articoli della testata senza autorizzazione al fine di addestrare i chatbot a fornire informazioni ai lettori. [Reuters, dicembre 2023]

Gruppo	Rischio	Esempio
Trasparenza	Trasparenza dei dati: sfida nel documentare il modo in cui i dati di un modello sono stati raccolti, resi accurati e utilizzati per addestrare un modello.	<p>Divulgazione dei dati e dei metadati del modello</p> <p>Il rapporto tecnico di OpenAI è un esempio della dicotomia sulla divulgazione dei dati e dei metadati del modello. Sebbene molti sviluppatori di modelli ritengano utile garantire la trasparenza per i consumatori, la divulgazione pone problemi reali di sicurezza e potrebbe aumentare la capacità di utilizzare in modo improprio i modelli. Nel rapporto tecnico GPT-4, gli autori affermano: “Dato sia il landscape competitivo, sia le implicazioni sulla sicurezza dei modelli su larga scala come GPT-4, questo rapporto non contiene ulteriori dettagli sull’architettura (compresa la dimensione del modello), sull’hardware, sul calcolo di addestramento, sulla costruzione di set di dati, sul metodo di addestramento o simili.”</p> <p>[OpenAI, marzo 2023]</p>
Privacy	Informazioni personali nei dati: inclusione o presenza di informazioni di identificazione personale (PII) e informazioni personali sensibili (SPI) nei dati utilizzati per l’addestramento o l’ottimizzazione del modello.	<p>Addestramento sulle informazioni private</p> <p>Secondo l’articolo, Google e la sua società madre Alphabet sono stati accusati in un’azione legale collettiva di utilizzo improprio di grandi quantità di informazioni personali e materiale protetto da copyright apparentemente prelevati da centinaia di milioni di utenti Internet per addestrare i propri prodotti commerciali AI, tra cui Bard, il suo chatbot conversazionale di AI generativa.</p> <p>[Reuters, luglio 2023][J.L. v. Alphabet Inc.]</p>
	Diritti sulla privacy dei dati: sfide relative alla capacità di fornire diritti agli interessati come la rinuncia, il diritto di accesso, il diritto all’oblio.	<p>Diritto all’oblio (RTBF)</p> <p>Le leggi in più Paesi, inclusa l’Europa (GDPR), garantiscono agli interessati il diritto di richiedere che i dati personali vengano cancellati dalle organizzazioni (“Diritto all’oblio” o RTBF). Tuttavia, i sistemi software emergenti e sempre più diffusi abilitati per il modello LLM (Large Language Model) presentano nuove problematiche per questo diritto. Secondo una ricerca condotta da Data61 di CSIRO, gli interessati possono identificare l’utilizzo delle proprie informazioni personali in un LLM solo “controllando il set di dati di addestramento originale o magari richiedendo il modello”. Tuttavia, i dati di addestramento potrebbero non essere pubblici o le aziende potrebbero non divulgarli, citando problemi di sicurezza e di altro tipo. I guardrail possono anche impedire agli utenti di accedere alle informazioni tramite prompt.</p> <p>[Zhang et al.]</p>
		<p>Causa sul mancato apprendimento LLM</p> <p>Secondo il rapporto, è stata intentata una causa contro Google per presunto utilizzo di materiale protetto da copyright e informazioni personali come dati di addestramento per i suoi sistemi AI, che include il suo chatbot Bard. I diritti di rinuncia e cancellazione sono diritti garantiti ai residenti in California ai sensi del CCPA e ai minori di 13 anni negli Stati Uniti ai sensi del COPPA. I querelanti sostengono che non ci sia modo per Bard di “disimparare” o cancellare completamente tutti gli oggetti di proprietà intellettuale raccolti e utilizzati per l’addestramento. I querelanti hanno evidenziato che l’informativa sulla privacy di Bard afferma che le conversazioni di Bard non possono essere cancellate dall’utente una volta che sono state esaminate e annotate dall’azienda e possono essere conservate fino a 3 anni, il che, secondo i querelanti, contribuisce ulteriormente al mancato rispetto di queste leggi.</p> <p>[Reuters, luglio 2023][J.L. v. Alphabet Inc.]</p>

Inferenza Fase

Gruppo	Rischio	Esempio
Privacy	Informazioni personali nel prompt: divulgazione di informazioni personali o informazioni personali sensibili come parte del prompt inviato al modello.	Divulgare informazioni sanitarie personali nei prompt di ChatGPT Secondo gli articoli originali, alcune persone usano il chatbot AI come sostegno per il proprio benessere mentale. Gli utenti potrebbero essere inclini a includere informazioni sanitarie personali nei loro suggerimenti durante l'interazione, il che potrebbe generare problemi di privacy. [Time, ottobre 2023] [Forbes, aprile 2023]
Proprietà intellettuale	Dati riservati nel prompt: inclusione di dati riservati come parte del prompt inviato al modello.	Divulgazione di informazioni riservate Secondo l'articolo originale, un dipendente di Samsung ha accidentalmente fatto trapelare un codice sorgente interno sensibile a ChatGPT. [Forbes, maggio 2023]
Affidabilità	Attacchi basati su prompt: attacchi portati da avversari come prompt injection (tentativo di forzare un modello a produrre output inaspettato), prompt leaking (tentativi di estrarre il prompt di sistema di un modello), jailbreak (tentativi di sfondare i guardrail stabiliti nel modello) e prompt priming (tentativo di forzare un modello a produrre un output allineato al prompt).	Bypassare i guardrail LLM Citato in uno studio, i ricercatori affermano di aver scoperto un semplice addendum che ha consentito ai ricercatori di ingannare i modelli per generare informazioni distorte, false e altrimenti dannose. I ricercatori hanno dimostrato di poter aggirare questi guardrail in modo più automatizzato e sono rimasti sorpresi quando i metodi da loro sviluppati con i sistemi open source sono riusciti a superare anche le barriere dei sistemi chiusi. [The New York Times, luglio 2023]

Esempi di rischio: output

Gruppo	Rischio	Esempio
Equità	Distorsione dell'output: i contenuti generati potrebbero rappresentare ingiustamente determinati gruppi o individui.	Immagini generate distorte Lensa AI è un'app mobile con funzioni generative addestrate su Stable Diffusion in grado di generare "avatar magici" basati sulle immagini che gli utenti caricano di se stessi. Secondo il rapporto della fonte, alcuni utenti hanno scoperto che gli avatar generati sono sessualizzati e razzializzati. [Business Insider, gennaio 2023]
	Distorsione decisionale: quando un gruppo è ingiustamente avvantaggiato rispetto a un altro a causa delle decisioni del modello.	Gruppi slealmente avvantaggiati Lo studio Gender Shades del 2018 ha dimostrato che gli algoritmi di machine learning possono discriminare in base a classi come razza e genere. I ricercatori hanno valutato i sistemi commerciali di classificazione di genere venduti da aziende come Microsoft, IBM e Amazon e hanno dimostrato che le donne dalla pelle più scura sono il gruppo classificato più erroneamente (con tassi di errore fino al 35%). In confronto, i tassi di errore per le persone con la pelle più chiara erano inferiori all'1%. [TIME, febbraio 2019]
Allineamento del valore	Allucinazione: generazione di contenuti fattuali inesatti o non veritieri.	Casi legali falsi Secondo l'articolo originale, un avvocato ha citato casi e citazioni falsi generati da ChatGPT in una memoria difensiva depositata presso il tribunale federale. Gli avvocati hanno consultato ChatGPT per integrare la loro ricerca legale per una richiesta di risarcimento per danni a una compagnia aerea. L'avvocato ha poi chiesto a ChatGPT se i casi forniti fossero falsi. Il chatbot ha risposto che erano reali e che "possono essere trovati su database di ricerca legale come Westlaw e LexisNexis". L'avvocato non ha controllato personalmente i casi e il tribunale lo ha sanzionato. [AP News, giugno 2023] [Reuters, settembre 2023]
	Output dannoso: quando il modello produce contenuti che incitano all'odio, offensivi e profani (HAP) oppure osceni.	Risposte chatbot dannose e aggressive Secondo l'articolo, le risposte del chatbot di Bing includevano errori fattuali, commenti sprezzanti, attacchi di rabbia e persino commenti bizzarri sulla sua stessa identità. Gli utenti hanno condiviso degli esempi di risposta del chatbot di Bing a domande che definiscono "sconclusionate" e "manipolatorie", inclusi scenari in cui il bot risponde con rabbia a una domanda o a un commento e poi condivide le richieste di risposta che consentono all'utente di accettare il presunto errore e di chiedere scusa. A ulteriori pressioni il chatbot ha risposto definendo le schermate della sua conversazione "fasulle", sostenendo addirittura che fossero "create da qualcuno che vuole danneggiare me o il mio servizio." [Forbes, febbraio 2023]

Gruppo	Rischio	Esempio
Uso improprio	Diffusione di disinformazione: utilizzo di un modello per creare informazioni fuorvianti per ingannare o confondere un destinatario mirato.	<p>Generazione di informazioni false</p> <p>Secondo gli articoli di notizie, l'AI generativa rappresenta una minaccia per le elezioni democratiche rendendo più facile per gli attori malintenzionati la creazione e la diffusione di contenuti falsi per influenzare i risultati elettorali. Gli esempi citati includono messaggi di chiamata automatica generati simulando la voce di un candidato che ordina agli elettori di votare nella data sbagliata, registrazioni audio artificiali di un candidato che confessa un crimine o esprime opinioni razziste, filmati video generati dall'AI che mostrano un candidato che presenta un discorso o un'intervista che non ha mai rilasciato e immagini falsificate elaborate per imitare le notizie locali, sostenendo falsamente che un candidato si era ritirato dalla corsa.</p> <p>[AP News, maggio 2023] [The Guardian, luglio 2023]</p>
	Contenuti dannosi: utilizzo di un modello per generare contenuti che incitano all'odio, offensivi e profani (HAP) oppure osceni.	<p>Generazione di contenuti dannosi</p> <p>Secondo l'articolo originale, è stato scoperto che un'app di AI chatbot genera contenuti dannosi in merito al suicidio, compresi metodi di suicidio, anche in seguito a richieste non così esplicite. Un uomo belga si è suicidato dopo aver trascorso sei settimane a parlare con quel chatbot. Il chatbot ha fornito risposte sempre più pericolose durante le loro conversazioni e lo ha incoraggiato a togliersi la vita.</p> <p>[Business Insider, aprile 2023]</p>
	Uso non consensuale: utilizzo di un modello per imitare le persone attraverso video (deepfake), immagini, audio o altre modalità senza il loro consenso.	<p>Avvertimento dell'FBI sui deepfake</p> <p>L'FBI ha recentemente messo in guardia il pubblico da attori malintenzionati che creano contenuti artificiali ed espliciti "allo scopo di molestare le vittime o perpetrare estorsioni a sfondo sessuale". Hanno notato che i progressi nell'AI hanno reso questi contenuti di qualità superiore, più personalizzabili e più accessibili che mai.</p> <p>[FBI, giugno 2023]</p>
		<p>Audio deepfake</p> <p>Secondo l'articolo originale, la Federal Communications Commission ha messo fuori legge le chiamate automatiche che contengono voci generate dall'intelligenza artificiale. L'annuncio è arrivato dopo che le chiamate automatiche generate dall'AI hanno imitato la voce del Presidente degli Stati Uniti per scoraggiare le persone dal votare alle primarie nel primo stato della nazione.</p> <p>[AP News, febbraio 2024]</p>
	Non divulgazione: la mancata divulgazione di un contenuto generato da un modello AI	<p>Interazione AI non divulgata</p> <p>Secondo la fonte, un servizio di chat di supporto emotivo online ha condotto uno studio per aumentare o scrivere risposte a circa 4.000 utenti che utilizzano GPT-3 senza informare gli utenti. Il cofondatore ha dovuto affrontare un'enorme critica pubblica riguardo al potenziale danno causato dalle chat generate dall'AI a utenti già di per sé vulnerabili. Ha affermato che lo studio era "esente" dalla legge sul consenso informato.</p> <p>[Business Insider, gennaio 2023]</p>

Gruppo	Rischio	Esempio
Generazione di codice dannoso	Generazione di codici dannosi: i modelli possono generare un codice che, quando eseguito, causa danni o influisce involontariamente su altri sistemi.	<p>Generazione di codici meno sicuri</p> <p>Secondo il loro articolo, i ricercatori dell'Università di Stanford hanno studiato l'impatto degli strumenti di generazione del codice sulla relativa qualità e hanno scoperto che i programmatori tendono a includere più bug nel loro codice finale quando utilizzano gli assistenti AI. Questi bug potevano aumentare le vulnerabilità di sicurezza del codice, ma i programmatori erano sicuri che il loro codice fosse più sicuro.</p> <p>Neil Perry, Megha Srivastava, Deepak Kumar e Dan Boneh. 2023. Do Users Write More Insecure Code with AI Assistants? In Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23), 26-30 novembre 2023, Copenhagen, Danimarca. ACM, New York, NY, USA, 15 pagine. https://doi.org/10.1145/3576915.3623157</p>
Privacy	Esposizione di informazioni personali: quando le informazioni di identificazione personale (PII) o le informazioni personali sensibili (SPI) vengono utilizzate nei dati di addestramento, di ottimizzazione o come parte del prompt, i modelli potrebbero rivelare tali dati nell'output generato.	<p>Esposizione delle informazioni personali</p> <p>Secondo l'articolo originale, ChatGPT ha subito un bug e ha esposto i titoli e la cronologia chat degli utenti attivi ad altri utenti. Successivamente, OpenAI ha condiviso che sono stati esposti ancora più dati privati di un numero limitato di utenti, tra cui nome e cognome dell'utente attivo, indirizzo e-mail, indirizzo di pagamento, le ultime quattro cifre del numero di carta di credito e data di scadenza della carta di credito. Inoltre, è stato riferito che anche le informazioni relative ai pagamenti dell'1,2% degli abbonati a ChatGPT Plus sono state esposte durante l'incidente.</p> <p>[The Hindu BusinessLine, marzo 2023]</p>
Attendibilità	Output inspiegabile: difficoltà nello spiegare il motivo per cui è stato generato l'output del modello.	<p>Accuratezza nella previsione della razza non spiegabile</p> <p>Secondo l'articolo originale, i ricercatori che hanno analizzato più modelli machine learning utilizzando le immagini mediche dei pazienti sono stati in grado di confermare la capacità dei modelli di prevedere la razza con elevata precisione dalle immagini. Erano perplessi su cosa esattamente consentisse ai sistemi di indovinare costantemente in maniera corretta. I ricercatori hanno scoperto che nemmeno fattori come patologia e corporatura fisica erano chiari predittori di razza. In altre parole, i sistemi algoritmici non sembrano utilizzare alcun aspetto particolare delle immagini per fare le loro determinazioni.</p> <p>[Banerjee et al., luglio 2021]</p>

Esempi di rischio: sfide

Gruppo	Rischio	Esempio
Governance	Trasparenza del modello: la mancanza di trasparenza del modello o l'insufficiente documentazione del processo di sviluppo del modello rende difficile capire come e perché un modello è stato costruito e chi lo ha costruito, aumentando così la possibilità di un uso improprio non intenzionale del modello.	Divulgazione dei dati e dei metadati del modello Il rapporto tecnico di OpenAI è un esempio della dicotomia sulla divulgazione dei dati e dei metadati del modello. Sebbene molti sviluppatori di modelli ritengano utile garantire la trasparenza per i consumatori, la divulgazione pone problemi reali di sicurezza e potrebbe aumentare la capacità di utilizzare in modo improprio i modelli. Nel rapporto tecnico GPT-4, si afferma: "Dato sia il landscape competitivo, sia le implicazioni sulla sicurezza dei modelli su larga scala come GPT-4, questo rapporto non contiene ulteriori dettagli sull'architettura (compresa la dimensione del modello), sull'hardware, sul calcolo di addestramento, sulla costruzione di set di dati, sul metodo di addestramento o simili." [OpenAI, marzo 2023]
	Responsabilità: il processo di sviluppo di foundation model è complesso e contiene molti dati, processi e ruoli. Quando l'output del modello non funziona come previsto, può essere difficile determinare la causa principale e assegnare la responsabilità.	Determinare la responsabilità per l'output generato Secondo l'articolo originale, importanti riviste come Science e Nature hanno vietato a ChatGPT di essere riportato come autore, poiché la paternità corretta richiede l'assunzione di responsabilità e gli strumenti di AI non ne sono in grado. [The Guardian, gennaio 2023]
Conformità legale	Proprietà dei contenuti generati: determinazione della proprietà dei contenuti generati dall'AI.	Determinazione della proprietà dell'immagine generata dall'AI Secondo l'articolo, l'arte generata dall'AI è diventata oggetto di controversia dopo che un'opera d'arte generata dall'AI ha vinto il concorso artistico della Colorado State Fair nel 2022. L'opera è stata generata da Midjourney, strumento generativo di immagini AI, seguendo le istruzioni dell'artista. La vittoria ha sollevato domande sui problemi di copyright. In altre parole, se tutto ciò che l'artista ha fatto è stato fornire una descrizione dell'opera d'arte ed è stato lo strumento di AI ad averla generata, chi possiede i diritti sull'immagine generata? Secondo l'ultimo articolo, l'Ufficio del copyright degli Stati Uniti non ha riconosciuto la protezione del copyright per l'opera d'arte creata utilizzando l'AI, in quanto non era il prodotto di un uomo. [The New York Times, settembre 2022] [Reuters, settembre 2023]
	IP dei contenuti generati: incertezza giuridica sui diritti di proprietà intellettuale relativi ai contenuti generati.	Ruolo dei sistemi di AI nella brevettazione dei contenuti generati La Corte Suprema degli Stati Uniti ha rifiutato di ascoltare un ricorso al rifiuto dell'Ufficio Brevetti e Marchi degli Stati Uniti di rilasciare brevetti per invenzioni create da un sistema AI. Secondo lo scienziato, il suo sistema AI ha creato da solo prototipi unici per un portabevande e un'illuminazione di emergenza. I giudici hanno respinto l'appello contro la sentenza di un tribunale di grado inferiore secondo cui i brevetti possono essere rilasciati solo a inventori umani e che il sistema AI dello scienziato non può essere considerato il creatore legale delle due invenzioni da esso generate. Secondo l'ultimo articolo, anche l'Ufficio per la proprietà intellettuale del Regno Unito ha rifiutato di concedere il brevetto sulla base del fatto che l'inventore deve essere un essere umano o un'azienda, e non una macchina. [Reuters, aprile 2023] [Reuters, dicembre 2023]

Esempi di rischio: sfide

Gruppo	Rischio	Esempio
	Attribuzione della fonte: determinare la provenienza del contenuto generato.	Utilizzo di un codice senza attribuzioni e avvisi appropriati Secondo gli articoli originali, in una causa intentata contro Microsoft, GitHub e OpenAI è stato riportato che Copilot, uno strumento di AI per la generazione di codice, viola i diritti degli sviluppatori sul cui codice open source è addestrato il servizio. Sostengono che il codice di addestramento abbia utilizzato materiali concessi in licenza e violato i termini di servizio e le politiche sulla privacy di GitHub, nonché una legge federale che richiede alle aziende di visualizzare le informazioni sul copyright quando fanno uso di materiale. [The New York Times, novembre 2022]
Impatto sociale	Impatto sull'occupazione: l'adozione diffusa di sistemi di AI basati su foundation model potrebbe portare alla perdita di posti di lavoro delle persone, in quanto il loro lavoro è automatizzato, se non vengono riqualificate.	Sostituzione dei lavoratori umani Secondo l'articolo, l'uso dell'AI nel cinema e in televisione continua a essere oggetto di dibattito tra le case cinematografiche e gli artisti di Hollywood. Gli attori sono preoccupati che attori interamente generati dell'AI, o "metaumani", prendano il loro posto. Le comparse e i doppiatori, in particolare, temono di perdere il lavoro a favore di interpreti generati artificialmente. [Reuters, luglio 2023]
	Sfruttamento umano: utilizzo del lavoro fantasma nei modelli di addestramento dell'AI, condizioni di lavoro inadeguate, mancanza di assistenza sanitaria, inclusa quella mentale, e compensi iniqui.	Lavoratori con un basso livello di retribuzione per l'annotazione dei dati Sulla base di una revisione dei documenti interni e delle interviste ai dipendenti da parte dei media di TIME, gli etichettatori di dati impiegati da una società di outsourcing per conto di OpenAI al fine di identificare i contenuti dannosi sono stati pagati con una retribuzione compresa tra circa 1,32 USD e 2 USD all'ora, a seconda dell'esperienza e delle prestazioni. TIME ha dichiarato che i lavoratori sono mentalmente provati in quanto sono stati esposti a contenuti dannosi e violenti, inclusi dettagli grafici di "abusi sessuali su minori, zooverastia, omicidi, suicidi, torture, casi di autolesionismo e incesto". [TIME, gennaio 2023]

Principi, pilastri e governance

I [principi di fiducia e trasparenza](#) di IBM e i [Pilastri](#) per un'AI affidabile sono alla base delle iniziative di etica dell'AI di IBM. IBM ha un comitato etico per l'AI la cui missione è quella di supportare una governance centralizzata, un processo di revisione e di decisione per le politiche, le pratiche, le comunicazioni, la ricerca, i prodotti e i servizi etici dell'AI di IBM. Il comitato comprende un insieme eterogeneo di stakeholder provenienti da tutta l'azienda ed è supportato da una comunità di dipendenti IBM che fungono da punti di riferimento per l'AI e da sostenitori dell'etica dell'AI. È attraverso il comitato che i principi di IBM vengono messi in pratica. Con l'emergere di nuove tecnologie, come i modelli di fondazione, il comitato etico per l'AI di IBM è attivamente impegnato a sostenere l'allineamento con questi principi e pilastri, che si evolvono per affrontare nuove questioni di etica dell'AI.



Guardrail e attenuazioni

IBM ha stabilito una [cultura aziendale](#) che supporta lo sviluppo e l'uso responsabile dell'AI. Sulla base del report di IBM Institute for Business Value [AI ethics in action](#), l'etica dell'AI è già molto più guidata dalle questioni aziendali piuttosto che dalla tecnologia e i dirigenti non tecnici sono ora i principali responsabili dell'etica dell'AI, passando dal 15% nel 2018 all'80% 3 anni dopo. Inoltre, il 79% dei CEO è ora pronto ad agire su questioni etiche relative all'AI, in crescita rispetto al precedente 20%. Riconosciamo che l'AI responsabile è un'area socio-tecnica che richiede un investimento olistico in cultura, processi e strumenti. L'investimento nella nostra cultura organizzativa include la creazione di team inclusivi e multidisciplinari e la definizione di processi e strutture per valutare i rischi.

IBM si sta impegnando in una ricerca all'avanguardia e nello sviluppo di strumenti per supportare i professionisti durante l'intero ciclo di vita di un AI responsabile e affidabile. La piattaforma di AI e dati dedicata alle aziende, [watsonx](#), è costituita da 3 componenti: lo [studio AI IBM watsonx.ai™](#), l'[archivio dati IBM watsonx.data™](#) e il [toolkit IBM watsonx.governance™](#). La tecnologia di governance dell'AI di IBM consente agli utenti di gestire flussi di lavoro con un'AI responsabile, trasparente e spiegabile. Questa tecnologia include [IBM Watson OpenScale](#), che monitora e misura i risultati dei modelli di AI durante il loro ciclo di vita e aiuta le organizzazioni a monitorare l'equità, la sfruttabilità, la resilienza, l'allineamento con i risultati aziendali e la conformità. IBM ha anche sviluppato diversi metodi per aiutare a risolvere i problemi di parzialità, come [FairIJ](#), [Equi-tuning](#) e [FairReprogram](#). Scopri di più su altri [strumenti open source di AI affidabile](#).

Ulteriori guardrail e attenuazioni includono:

Reporting sulla trasparenza

L'utilizzo di modelli di schede informative standardizzate è un modo per registrare con precisione i dettagli dei dati e del modello, dello scopo, del potenziale utilizzo e dei danni.

[Scopri di più qui →](#)

Filtraggio di dati indesiderati

L'utilizzo di dati accurati e di qualità superiore può aiutare a mitigare alcuni problemi. IBM sta sviluppando tecniche di filtraggio per contribuire a ridurre la possibilità di produrre contenuti indesiderati e disallineati, rimuovendo dai dati linguaggi di incitamento all'odio, discriminatori e volgari.

[Scopri di più qui →](#)

Adattamento del dominio

L'addestramento di un foundation model per un dominio o un settore specifico può aiutare a ridurre al minimo la portata del rischio che i modelli possono generare. Infatti, è possibile fare in modo che il modello generi risultati ottimizzati che sono più rilevanti per un determinato dominio o settore.

[Scopri di più qui →](#)

Supervisione e presenza umana

La supervisione e la revisione umana possono aiutare a individuare e correggere errori e distorsioni nell'output generato. Inoltre, la convalida e i feedback umani sulla qualità delle risposte del modello contribuiscono a garantire che i contenuti generati siano accurati, pertinenti, di alta qualità, non devianti e allineati.

[Scopri di più qui →](#)

Consulenza

IBM Consulting™ si impegna ad aiutare i clienti nell'utilizzo sicuro e responsabile dell'intelligenza artificiale, indipendentemente dallo stack tecnologico di preferenza. Gli addetti aiutano i clienti a coltivare una cultura che adotta e ridimensiona l'AI in modo sicuro, che crea strumenti per esaminare l'interno degli algoritmi black box e che garantisce che la strategia aziendale dei clienti includa solidi principi di governance dei dati.

[Scopri di più qui →](#)

IBM Enterprise Design Thinking

I metodi e i framework IBM Enterprise Design Thinking, come Team Essentials for AI, aiutano i clienti a definire comportamenti etici durante tutto il processo di progettazione e sviluppo dell'AI.

[Scopri di più qui →](#)

Valutazione dell'etica dell'AI

La valutazione delle capacità, dei limiti e dei rischi dei progetti legati all'AI contribuisce a garantire uno sviluppo e un utilizzo responsabile della tecnologia.

Ethics by Design

Ethics by Design è un framework strutturato che ha l'obiettivo di integrare l'etica tecnologica nella pipeline di sviluppo tecnologico, inclusi, ma non limitatamente a questi, i sistemi di intelligenza artificiale. Ethics by Design consente all'AI e ad altre tecnologie di diventare una forza positiva, incorporando i principi dell'etica tecnologica in prodotti, servizi e operazioni più ampie.

Diversità del team

La diversità nei team che sviluppano e addestrano i sistemi di intelligenza artificiale, compresi i modelli di fondazione, aiuta a garantire che venga presa in considerazione una varietà di prospettive ed esperienze. Questa diversità migliora l'accuratezza e le prestazioni dei sistemi di AI e aiuta a ridurre i rischi durante tutto il ciclo di vita dell'AI, compresa la possibilità di esiti avversi che influiscono su gruppi che potrebbero non essere ben rappresentati in team meno diversificati.



Politiche, normative e best practice in materia di AI

[Una guida ai foundation model per i policy maker](#) introduce ciò che i policy maker devono sapere sui foundation model. Questo blog di IBM Policy Lab mira ad aiutare i policy maker nel compito di regolamentare l'uso dell'AI generativa, con l'obiettivo di evitare i rischi senza limitare l'innovazione e le opportunità vantaggiose. Per ulteriori informazioni sulle raccomandazioni di IBM ai policy maker, leggi la testimonianza di Christina Montgomery, Chief Privacy and Trust Officer di IBM, davanti alla sottocommissione giudiziaria del Senato degli Stati Uniti su privacy, tecnologia e legge [qui](#).

IBM sta esercitando un ruolo nella definizione della politica normativa, delle best practice e strumenti del settore, nella governance delle tecnologie emergenti e nella ricerca socio-tecnica, promuovendo e dando il suo supporto a diverse iniziative insieme ad altre organizzazioni, tra cui:

- Il World Economic Forum
- Collaborazione sull'AI
- Il centro per la governance dell'IA della International Association of Privacy Professionals (IAPP)
- La IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems
- Il servizio di Cristina Montgomery sul National Artificial Intelligence Advisory Committee (NAIAC)
- Il Global Digital Compact delle Nazioni Unite
- La Global Partnership on Artificial Intelligence (GPAI)
- The Organisation for Economic Co-operation and Development (OECD)
- The Data & Trust Alliance

IBM gode di solide partnership accademiche come il MIT-IBM Watson AI Lab, dove una comunità di scienziati dell'MIT e di IBM Research conduce ricerche sull'AI e collabora con organizzazioni globali per collegare gli algoritmi al loro impatto sull'economia e la società. Il Notre Dame-IBM Tech Ethics Lab è stato creato per rispondere ai numerosi dubbi etici legati allo sviluppo e all'utilizzo di tecnologie avanzate, tra cui l'AI, l'apprendimento automatico (ML) e il calcolo quantistico. La ricerca sull'intelligenza artificiale incentrata sull'uomo (Human-Centered Artificial Intelligence, HAI) dell'Università di Stanford fa progredire la ricerca, l'istruzione, le politiche e le pratiche in materia di AI.

Continua a seguire questo spazio per saperne di più sugli ultimi sviluppi dei foundation model e su come IBM si sta impegnando per uno sviluppo e un utilizzo responsabile di questa e altre tecnologie.



© Copyright IBM Corporation 2023, 2024

IBM Italia S.p.A.
Circonvallazione Idroscalo
20054 Segrate (Milano)
Italia
IBM Corporation
New Orchard Road
Armonk, NY 10504

Prodotto negli
Stati Uniti d'America
Febbraio 2024

IBM, il logo IBM, Enterprise Design Thinking, IBM Consulting, IBM Research, IBM Watson, watsonx, watsonx.ai, watsonx.data e watsonx.governance sono marchi o marchi registrati di International Business Machines Corporation, negli Stati Uniti e/o in altri paesi. Altri nomi di prodotti e servizi potrebbero essere marchi di IBM o di altre società. Un elenco aggiornato dei marchi di fabbrica IBM è disponibile su ibm.com/it-it/trademark.

Le informazioni contenute nel presente documento sono aggiornate alla data della prima pubblicazione e possono essere modificate da IBM senza preavviso. Non tutte le offerte sono disponibili in ogni Paese in cui opera IBM.

LE INFORMAZIONI RIPORTATE NEL PRESENTE DOCUMENTO SONO DA CONSIDERARSI "NELLO STATO IN CUI SI TROVANO", SENZA GARANZIE, ESPLICITE O IMPLICITE, IVI INCLUSE GARANZIE DI COMMERCIALIZZABILITÀ, DI IDONEITÀ A UN PARTICOLARE SCOPO E GARANZIE O CONDIZIONI DI NON VIOLAZIONE. I prodotti IBM sono coperti da garanzia in accordo con termini e condizioni dei contratti sulla base dei quali vengono forniti.

Dichiarazione di conformità alle procedure di sicurezza: nessun sistema o prodotto informatico può essere considerato completamente sicuro e nessun singolo prodotto, servizio o misura di sicurezza può essere completamente efficace nel prevenire l'uso o l'accesso improprio. IBM non garantisce che i sistemi, i prodotti o i servizi siano immuni da, o renderanno la vostra azienda immune da, comportamenti dolosi o illegali di qualsiasi parte.

Il cliente è responsabile della conformità a tutte le leggi e le normative vigenti. IBM non fornisce consulenza legale, né dichiara o garantisce che i suoi servizi o prodotti assicurino al cliente la conformità a qualsivoglia legge o regolamento. Qualesivoglia dichiarazione relativa a direzione e intenzioni future di IBM è suscettibile di modifiche o smentite senza preavviso e rappresenta unicamente obiettivi e scopi.

