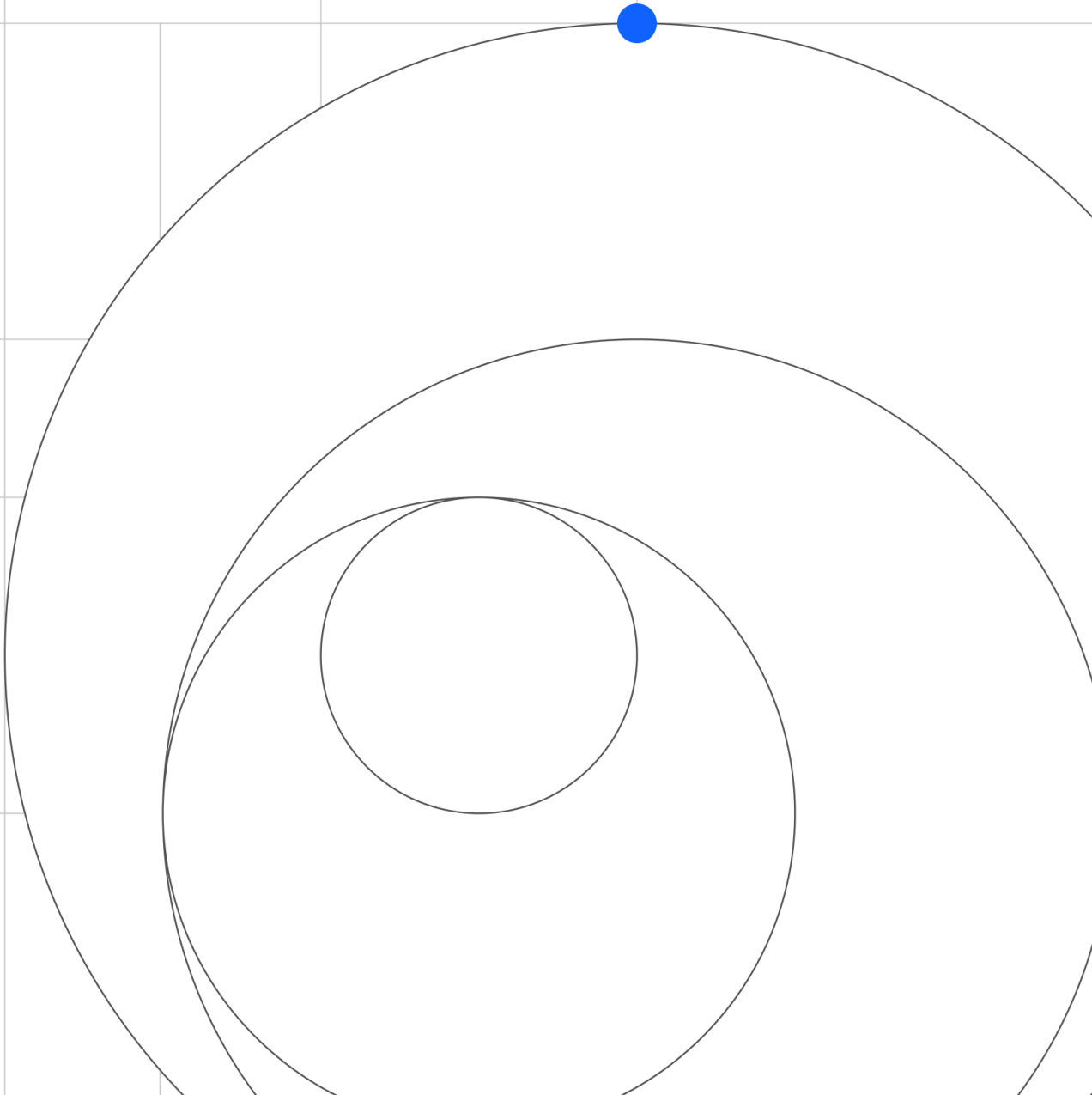
Modelos de base: oportunidades, riscos e mitigações



Atribuição

Com gratidão aos patrocinadores executivos do grupo de trabalho da ética em IA, Christina Montgomery e Francesca Rossi, e às contribuições dos membros do grupo de trabalho 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.

Índice

04
Executivo Resumo

05
Introdução

06
Benefícios dos modelos de base

08
Riscos dos modelos de base

16
Risco Exemplos

24
Princípios, pilares e governança

25
Proteções e mitigações

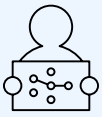
27
Políticas, regulamentações e melhores práticas de IA Exemplos

Resumo executivo

A ascensão dos modelos de base oferece às empresas novas e empolgantes possibilidades, mas também levanta questões novas e amplas sobre design, desenvolvimento, implementação e uso ético. Segundo uma recente pesquisa sobre IA generativa do IBM Institute for Business Value, as organizações já estão manifestando preocupações sobre questões relacionadas à confiança, especificamente como barreiras para investimentos. Suas principais preocupações são cibersegurança (57%), privacidade (51%) e precisão (47%). Muitas organizações estavam levando essas preocupações a sério antes da *'consumerização'* da IA generativa, expressando sua intenção de investir pelo menos 40% mais em ética de IA nos próximos três anos. A conscientização sobre riscos e possíveis maneiras de mitigá-los é o primeiro passo crucial para a criação de sistemas de IA confiáveis.

Neste documento:

Exploraremos as vantagens dos modelos de base, incluindo sua capacidade de realizar tarefas desafiadoras, potencial para acelerar a adoção de IA, habilidade de aumentar a produtividade e os benefícios econômicos que eles proporcionam.



Discutiremos as três categorias de risco, incluindo riscos conhecidos de formas anteriores de IA, riscos conhecidos amplificados por modelos de base e riscos emergentes intrínsecos aos recursos generativos dos modelos de base.

Abordaremos os princípios, os pilares e o controle que formam a base das iniciativas éticas de IA da IBM e sugeriremos barreiras para a mitigação de riscos.

Introdução

À medida que o uso de IA continua se expandindo, os grandes e complexos modelos de IA estão fornecendo resultados promissores de desempenho, bem como resolvendo alguns dos problemas mais desafiadores da sociedade. No entanto, criar grandes conjuntos de dados de treinamento e modelos complexos para cada aplicativo de IA pode ser extremamente difícil para as empresas. Modelos de base fornecem um caminho para alcançar o melhor dos dois mundos: desenvolver modelos de última geração poderosos e reutilizá-los diretamente ou aplicar métodos de ajuste para implementar uma variedade de casos de uso, em vez de treinar novos modelos para cada caso de uso. Por exemplo, a IBM Research desenvolveu modelos de base para inspeção visual. Esses modelos de base aprendem a representação geral de superfícies e corredores de concreto e podem ser ajustados ainda mais para casos de uso específicos, como detecção de rachaduras ou inspeção de defeitos com dados menos rotulados.

A IBM define um *modelo de base* como um modelo de IA que pode ser adaptado a uma ampla gama de tarefas de recebimento de dados. Os modelos de base normalmente são modelos generativos de grande escala treinados em dados não rotulados usando autossupervisão. Como modelos de grande escala, os modelos de base podem incluir bilhões de parâmetros.

A IBM é uma empresa de nuvem híbrida e IA com vasta reputação como administradora de dados responsável e comprometida com a ética em IA. Usando a capacidade de nossas equipes de pesquisa, produto e consultoria, juntamente com parceiros externos, como a Hugging Face, ajudamos a trazer o poder dos modelos de base para nossos clientes e a criar IAs confiáveis em qualquer empresa. A IBM também continua investindo na criação de novas plataformas, como a IA IBM watsonx e plataformas e tecnologias de dados, para projetar e desenvolver modelos de IA para se comportar de maneira auditável e confiável.

Este documento descreve o ponto de vista da IBM sobre a ética dos modelos de base. É a primeira versão, e as versões futuras expandirão vários aspectos da abordagem ética do modelo de base da IBM. Esperamos que este documento seja útil para todos os stakeholders no desenvolvimento, implementação e uso do modelo de base de forma responsável.

Benefícios dos modelos de base

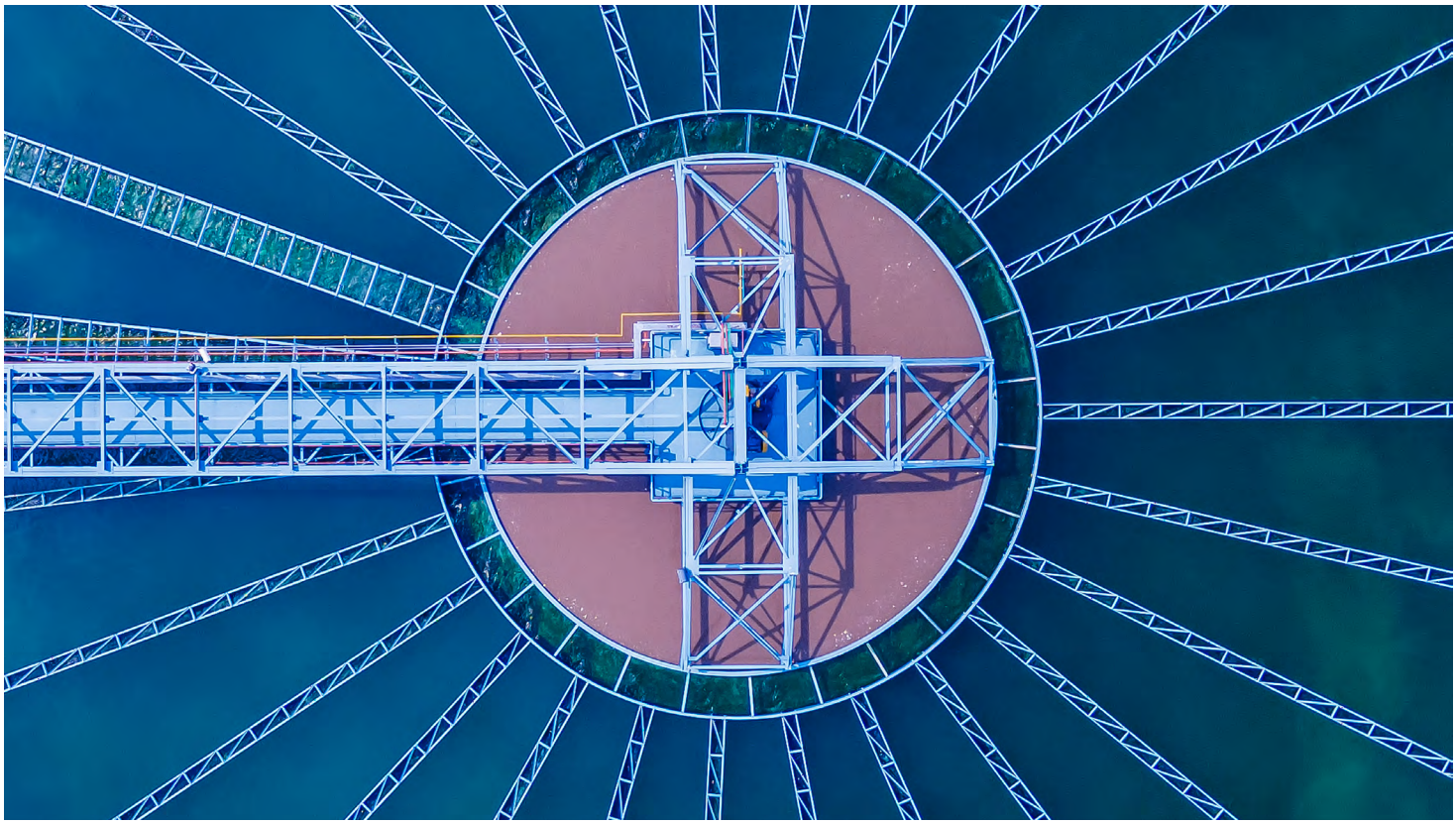
Os modelos de base podem melhorar significativamente o processo de desenvolvimento de sistemas de IA e auxiliar no avanço da IA da fase de exploração para a adoção nas empresas. Seus benefícios incluem:

Realizar tarefas complexas
Modelos de base mostram um aumento significativo no desempenho na resolução de problemas complexos e difíceis. Por exemplo, o modelo de base geoespacial da colaboração IBM e NASA foi projetado para converter os dados de satélite da NASA em mapas de desastres naturais, como inundações e outras mudanças de cenário. O modelo também pode ser usado para ajudar a revelar o passado do nosso planeta; estimar riscos para culturas, empresas ou infraestruturas devido ao clima severo; desenvolver estratégias para se adaptar às mudanças climáticas; e auxiliar no agronegócio. O modelo está planejado para ser disponibilizado previamente aos clientes IBM por meio do IBM Environmental Intelligence Suite.

Para ilustrar, o MoLFormer-XL da IBM é um modelo de base que é capaz de inferir a estrutura de moléculas a partir de representações simples, tornando mais fácil a aprendizagem de várias tarefas de recebimento de dados, como prever as propriedades físicas e quânticas de uma molécula, identificar moléculas semelhantes, rastrear moléculas já aprovadas para novos casos de uso e descobrir novas moléculas. Moderna e IBM estão explorando formas de usar o MoLExer para ajudar a prever propriedades das moléculas e entender as características de possíveis medicamentos de mRNA.

Maior produtividade
A natureza generativa dos modelos de base amplia o número de áreas em que a IA pode ser usada em uma empresa para ajudar a melhorar a produtividade, automatizando tarefas rotineiras e tediosas e permitindo que os usuários dediquem mais tempo ao trabalho criativo e inovador. Por exemplo, o IBM Watsonx Code Assistant, desenvolvido com modelos de base, possibilita que desenvolvedores, independentemente do nível de experiência, escrevam códigos usando recomendações geradas por IA.

Time to value mais rápido
Modelos de base geralmente são treinados com dados não rotulados, que estão mais disponíveis em grandes quantidades do que dados rotulados. Uma vez treinados, os modelos de base podem ser usados diretamente ou após serem ajustados para aplicativos de recebimento de dados, usando uma pequena quantidade de dados rotulados especializados, que podem diminuir a criação do time to value.



Utilize diversas modalidades de dados
Os modelos de base podem ser treinados usando diversas modalidades de dados, como língua natural, texto, imagem e áudio. Eles também podem ser aplicados a tarefas que exigem diferentes tipos de dados, como dados de séries temporais, dados geoespaciais, dados tabulares, dados semiestruturados e dados de modalidade mista, como texto combinado com imagens.

Despesas amortizadas
Embora o custo inicial do treinamento de um modelo de base seja significativamente maior do que o treinamento de um modelo de IA tradicional, o custo adicional para aplicá-lo em uma nova tarefa é consideravelmente menor. O uso de modelos de base pré-treinados poderia ajudar a eliminar a necessidade de que as empresas façam investimentos substanciais para treinar modelos de base e explorar suas novas capacidades. Para uma empresa, a confiabilidade dos modelos, a eficiência energética, o desempenho, a portabilidade e a capacidade de usar dados corporativos de forma eficaz e segura são fundamentais.

A IBM permite que as empresas criem e detenham o valor de modelos de base para seus negócios, trazendo as melhores inovações da comunidade de IA aberta e global, operando de forma eficiente em ambientes de computação híbrida, ajudando a mitigar riscos e controlando rigorosamente a IA.

Riscos dos modelos de base

Como todas as tecnologias que avançam rapidamente, os modelos de base oferecem riscos e benefícios. Alguns são riscos legais, como restrições à movimentação ou uso de dados, e precisam ser cuidadosamente avaliados de acordo com a legislação atual e em evolução. Outros riscos têm uma natureza ética e devem ser considerados cuidadosamente para que a tecnologia tenha um impacto positivo. Em geral, os riscos de IA levantam questões sociotécnicas e devem ser abordados e mitigados por meio de métodos sociotécnicos, incluindo ferramentas de software, processos de avaliação de risco, frameworks de ética em IA, mecanismos de controle, consultas multistakeholder, padrões e regulamentação. Iremos listar os riscos considerando as seguintes 3 categorias:

1. **Tradicional.** Riscos conhecidos de formas anteriores ou anteriores de sistemas de IA
2. **Amplificados.** Riscos conhecidos, mas agora intensificados devido às características intrínsecas dos modelos de base, principalmente seus recursos generativos inerentes
3. **Novo.** Riscos emergentes intrínsecos aos modelos de base e suas capacidades generativas inerentes

Também estruturamos a lista de riscos em relação a se estão principalmente associados ao conteúdo fornecido ao modelo base, o input, ou ao conteúdo gerado por ele, o output, ou se estão relacionados a desafios adicionais.

1. Riscos associados à entrada

Fase de treinamento e ajuste

Grupo	Risco	Por que isso é uma preocupação?	Indicador
Justiça	Viés de dados: viés histórico, representacional e social presente nos dados usados para treinar e fazer o ajuste fino do modelo.	Treinar um sistema de IA com dados enviesados, como viés histórico ou representacional, pode resultar em outputs enviesados ou distorcidos que podem representar injustamente ou discriminar certos grupos ou indivíduos. Além dos impactos negativos na sociedade, entidades comerciais podem enfrentar consequências legais, interrupção das operações ou danos à reputação decorrentes dos resultados enviesados do modelo.	Amplificado
Robustez	Envenenamento de dados: um tipo de ataque adversarial no qual um adversário ou agente interno malicioso injeta intencionalmente amostras corrompidas, falsas, enganosas ou incorretas no conjunto de dados de treinamento ou ajuste fino.	O envenenamento de dados pode tornar o modelo sensível a um padrão de dados malicioso e produzir o output desejado pelo adversário. Isso pode criar um risco de segurança onde adversários podem manipular o comportamento do modelo em seu próprio benefício. Além de produzir resultados não intencionais e potencialmente maliciosos, uma divergência do modelo causada por envenenamento de dados pode resultar em entidades comerciais enfrentando consequências legais, interrupção das operações ou danos à reputação.	Tradicional
Alinhamento de valor	Curadoria de dados: quando os dados de treinamento ou ajuste são coletados ou preparados de forma inadequada.	Uma curadoria de dados inadequada pode afetar adversamente como um modelo é treinado, resultando em um modelo que não se comporta de acordo com os valores pretendidos. Exemplos de uma curadoria de dados inadequada podem incluir erros de rotulagem ou anotação nos dados usados para treinar ou ajustar o modelo. Corrigir problemas após o treinamento e a implementação do modelo pode ser insuficiente para garantir um comportamento adequado. Um comportamento inadequado do modelo pode resultar em entidades comerciais enfrentando consequências legais, interrupções nas operações ou danos à reputação.	Amplificado
	Retreinamento baseado em downstream: usando de outputs indesejáveis (imprecisos, inadequados, conteúdo do usuário, etc.) de aplicações downstream para fins de retreinamento.	O reaproveitamento de output downstream para treinar novamente um modelo sem implementar a verificação humana adequada aumenta as chances de que outputs indesejáveis sejam incorporados aos dados de treinamento ou ajuste do modelo, possivelmente gerando outputs ainda mais indesejáveis. Comportamento inadequado do modelo pode resultar em entidades empresariais enfrentando consequências legais ou danos à reputação. Não cumprir com as leis de transferência de dados pode resultar em multas e outras consequências legais.	Novo
Leis de dados	Transferência de dados: leis e outras restrições podem limitar ou proibir a transferência de dados.	Restrições à transferência de dados podem afetar a disponibilidade dos dados necessários para treinar um modelo de IA e podem resultar em dados mal representados. Além do impacto na disponibilidade de dados, o não cumprimento das leis e regulamentações de transferência de dados pode resultar em multas e outras consequências legais.	Tradicional
	Uso de dados: leis e outras restrições podem limitar ou proibir o uso de alguns dados para casos de uso específicos de IA.	O não cumprimento das leis e regulamentações de uso de dados pode resultar em multas e outras consequências legais.	Tradicional
	Aquisição de dados: leis e outras regulamentações podem limitar a coleta de certos tipos de dados para casos de uso específicos de IA.	O não cumprimento das leis e regulamentações da aquisição de dados pode resultar em multas e outras consequências legais.	Amplificado

Grupo	Risco	Por que isso é uma preocupação?	Indicador
Propriedade intelectual	Direitos de uso de dados: termos de serviço, leis de direitos autorais, conformidade com licenças ou outras questões de propriedade intelectual podem restringir a capacidade de usar certos dados para a construção de modelos.	As leis e regulamentações referentes ao uso de dados para treinar IA são instáveis e podem variar de país para país, o que cria desafios no desenvolvimento de modelos. Se o uso de dados violar regras ou restrições, as entidades empresariais podem enfrentar multas, danos à reputação, interrupção das operações e outras consequências legais.	Amplificado
Transparência	Transparência de dados: desafio em documentar como os dados de um modelo foram coletados, curados e utilizados para treiná-lo.	A transparência dos dados é importante para a conformidade legal e ética da IA. A falta de informações limita a capacidade de avaliar os riscos associados aos dados. A falta de requisitos padronizados pode limitar a divulgação, pois as organizações protegem segredos comerciais e tentam evitar que outros copiem seus modelos.	Amplificado
	Procedência dos dados: desafio em padronizar e estabelecer métodos para verificar de onde os dados vieram.	Nem todas as fontes de dados são confiáveis. Os dados podem ter sido coletados, manipulados ou falsificados de forma antiética. O uso de dados não confiáveis pode resultar em comportamentos indesejáveis no modelo. As entidades empresariais podem enfrentar multas, danos à reputação, interrupção das operações e outras consequências legais.	Amplificado
Privacidade	Informações pessoais nos dados: inclusão ou presença de informações pessoalmente identificáveis (PII) e informações pessoais sensíveis (SPI) nos dados usados para treinar ou ajustar o modelo.	Se não desenvolvido adequadamente para proteger dados sensíveis, o modelo pode expor informações pessoais no output gerado. Além disso, dados pessoais ou sensíveis devem ser revisados e tratados de acordo com as leis e regulamentações de privacidade. As entidades empresariais podem enfrentar multas, danos à reputação, interrupção das operações e outras consequências legais se forem encontradas em violação.	Tradicional
	Reidentificação: mesmo com a remoção de informações pessoalmente identificáveis (PII) e informações pessoais sensíveis (SPI) dos dados, ainda pode ser possível identificar pessoas devido a outros recursos disponíveis nos dados.	Os dados que podem revelar informações pessoais ou sensíveis devem ser revisados com respeito às leis e regulamentações de privacidade, pois as entidades comerciais podem enfrentar multas, danos à reputação, interrupção das operações e outras consequências legais se forem consideradas em violação.	Tradicional
	Direitos de privacidade de dados: desafios relacionados à capacidade de fornecer direitos do titular dos dados, como opção de exclusão, direito de acesso e direito ao esquecimento.	A identificação ou uso inadequado de dados pode resultar em violação das leis de privacidade. O uso inadequado ou um pedido de remoção de dados poderia obrigar as organizações a reconfigurar o modelo, o que é caro. Além disso, as entidades empresariais podem enfrentar multas, danos à reputação, interrupção das operações e outras consequências legais se não cumprirem as regras e regulamentações de privacidade de dados.	Amplificado
	Consentimento informado: dados coletados para treinar modelos de IA sem o consentimento informado do proprietário, mesmo quando legalmente permitido.	Em algumas circunstâncias, pode ser antiético coletar e usar dados sem o consentimento da pessoa. Existem também possíveis riscos reputacionais associados a esse tipo de uso.	Tradicional

Inferência Fase

Grupo	Risco	Por que isso é uma preocupação?	Indicador
Privacidade	Informações pessoais no prompt: divulgar informações pessoais ou informações pessoais sensíveis como parte do prompt solicitação enviada ao modelo.	Os dados do prompt podem ser armazenados ou posteriormente utilizados para outros fins, como avaliação e retreinamento do modelo. Esses tipos de dados devem ser revisados com respeito às leis e regulamentações de privacidade. Sem um armazenamento e uso adequados dos dados, as entidades empresariais podem enfrentar multas, danos à reputação, interrupção das operações e outras consequências legais.	Novo
Propriedade intelectual	Informações de IP no prompt: divulgação de informações de direitos autorais ou outras informações de propriedade intelectual como parte do prompt enviado ao modelo.	Os dados do prompt podem ser armazenados ou posteriormente utilizados para outros fins, como avaliação e retreinamento do modelo. Esses tipos de dados devem ser revisados com respeito às leis e regulamentações de propriedade intelectual. Sem um armazenamento e uso adequados dos dados, as entidades empresariais podem enfrentar multas, danos à reputação, interrupção das operações e outras consequências legais.	Novo
	Dados confidenciais no prompt: inclusão de dados confidenciais como parte do prompt enviado ao modelo.	Se não for desenvolvido adequadamente para proteger dados confidenciais, o modelo pode expor informações confidenciais ou propriedade intelectual no output gerado. Além disso, informações confidenciais dos usuários finais podem ser coletadas e armazenadas inadvertidamente.	Novo
Robustez	Ataque de evasão: tentativa de fazer com que um modelo produza outputs incorretos perturbando os dados enviados ao modelo treinado.	Os ataques de evasão alteram o comportamento do modelo, geralmente para beneficiar o atacante. Se os resultados de output não forem devidamente considerados, as entidades empresariais podem enfrentar multas, danos à reputação, interrupção das operações e outras consequências legais.	Amplificado
	Ataques baseados em prompt: ataques adversos, como injeção de prompt (tentativa de forçar um modelo a produzir um output inesperado), vazamento de prompt (tentativas de extrair o prompt do sistema de um modelo), desbloqueio (tentativas de romper as proteções estabelecidas no modelo), e preparação de prompt (tentativa de forçar um modelo a produzir um output alinhado ao prompt).	Dependendo do conteúdo revelado, as entidades empresariais podem enfrentar multas, danos à reputação, interrupção das operações e outras consequências legais.	Novo

2. Riscos associados à saída

Grupo	Risco	Por que isso é uma preocupação?	Indicador
Justiça	Viés de output: o conteúdo gerado pode representar injustamente certos grupos ou indivíduos.	O viés pode prejudicar os usuários dos modelos de IA e amplificar comportamentos discriminatórios existentes. As entidades empresariais podem enfrentar danos à reputação, interrupção das operações e outras consequências.	Novo
	Viés de decisão: quando um grupo é injustamente favorecido em relação a outro devido aos efeitos das decisões tomadas por humanos usando o output do modelo.	O viés pode prejudicar as pessoas afetadas pelas decisões do modelo. As entidades empresariais podem enfrentar multas, danos à reputação, interrupção das operações e outras consequências legais.	Tradicional
Propriedade intelectual	Violação de direitos autorais: quando um modelo gera conteúdo que é muito semelhante ou idêntico a uma obra existente protegida por direitos autorais ou abrangida por um acordo de licença de código aberto.	As leis e regulamentações referentes ao uso de conteúdo que se assemelha ou é muito semelhante a outros dados protegidos por direitos autorais são amplamente indefinidos e podem variar de país para país, o que representa desafios na determinação e implementação da conformidade. As entidades empresariais podem enfrentar multas, danos à reputação, interrupção das operações e outras consequências legais.	Novo
Alinhamento de valor	Alucinação: geração de conteúdo factualmente impreciso ou não verdadeiro.	Outputs falsos podem induzir os usuários ao erro e serem incorporados em artefatos posteriores, propagando ainda mais a desinformação. Isso pode prejudicar tanto os proprietários quanto os usuários dos modelos de IA. Também, as entidades empresariais podem enfrentar multas, danos à reputação, interrupção das operações e outras consequências legais.	Novo
	Outputs tóxicos: quando o modelo produz conteúdo odioso, abusivo e profano (HAP) ou obsceno.	Conteúdo odioso, abusivo e profano (HAP) ou obsceno pode impactar adversamente e prejudicar as pessoas que interagem com o modelo. Também, as entidades empresariais podem enfrentar multas, danos à reputação, interrupção das operações e outras consequências legais.	Novo
	Conselhos perigosos: quando um modelo fornece conselhos sem ter informações suficientes, resultando em possíveis perigos se o conselho for seguido.	Uma pessoa pode agir com base em conselhos incompletos ou preocupar-se com uma situação que não se aplica a ela devido à natureza supergeneralizada do conteúdo gerado.	Novo
Uso indevido	Disseminação de desinformação: utilização de um modelo para criar informações enganosas ou falsas com o intuito de enganar ou influenciar um público-alvo.	Espalhar desinformação pode afetar a capacidade de uma pessoa de tomar decisões informadas. As entidades empresariais podem enfrentar multas, danos à reputação, interrupção das operações e outras consequências legais.	Novo
	Toxicidade: utilizar um modelo para gerar conteúdo odioso, abusivo e profano (HAP) ou obsceno.	Conteúdo tóxico pode ter um impacto negativo no bem-estar de seus destinatários. As entidades empresariais podem enfrentar multas, danos à reputação, interrupção das operações e outras consequências legais.	Novo
	Uso não consensual: utilizar um modelo para imitar pessoas por meio de vídeo (deepfakes), imagens, áudio ou outras modalidades sem o consentimento delas.	Deepfakes podem disseminar desinformação sobre uma pessoa, possivelmente resultando em impactos negativos na reputação da pessoa. As entidades empresariais podem enfrentar multas, danos à reputação, interrupção das operações e outras consequências legais.	Amplificado

Grupo	Risco	Por que isso é uma preocupação?	Indicador
	Uso perigoso: utilizar um modelo com a única intenção de prejudicar pessoas.	As entidades empresariais podem enfrentar multas, danos à reputação, interrupção das operações e outras consequências legais.	Novo
	Não divulgação: não revelar que o conteúdo é gerado por um modelo de IA.	A omissão do conteúdo produzido por IA pode ser interpretada como enganosa, levando a uma diminuição da confiança. A intenção de enganar pode resultar na redução da capacidade de ação humana, em multas, danos à reputação e outras consequências legais.	Novo
	Uso inadequado: utilizar um modelo para um fim para o qual o modelo não foi projetado.	Reutilizar um modelo sem compreender seus dados originais, intenção de design e objetivos pode resultar em comportamentos inesperados e indesejados do modelo.	Amplificado
Geração de código prejudicial	Geração de código prejudicial: modelos podem gerar código que, quando executado, causa danos ou afeta inadvertidamente outros sistemas.	A execução de código prejudicial pode abrir vulnerabilidades nos sistemas de TI. As entidades empresariais podem enfrentar multas, danos à reputação, interrupção das operações e outras consequências legais.	Novo
Confiança equivocada	Excesso/falta de confiança: quando uma pessoa deposita confiança em excesso ou em falta na orientação de um modelo de IA.	Em tarefas onde os humanos baseiam suas escolhas em sugestões da IA, uma confiança excessiva ou insuficiente pode levar a decisões inadequadas devido à confiança equivocada no sistema de IA, com consequências negativas que aumentam com a importância da decisão. Decisões ruins podem prejudicar as pessoas e podem resultar em prejuízos financeiros, danos à reputação, interrupção das operações e outras consequências legais para as entidades comerciais.	Amplificado
Privacidade	Expor informações pessoais: quando informações pessoalmente identificáveis (PII) ou informações pessoais sensíveis (SPI) são utilizadas nos dados de treinamento, dados de ajuste fino ou como parte do prompt, os modelos podem revelar esses dados no output gerado.	Compartilhar informações pessoalmente identificáveis das pessoas afeta seus direitos e as torna mais vulneráveis. Além disso, os dados dos outputs devem ser revisados em conformidade com as leis e regulamentações de privacidade, pois as entidades comerciais podem enfrentar multas, danos à reputação, interrupção das operações e outras consequências legais se forem encontradas em violação das leis ou regulamentações de privacidade ou uso de dados.	Novo
Explicabilidade	Output inexplicável: desafios em explicar por que o output do modelo foi gerado.	Os modelos de base são baseados em arquiteturas complexas de deep learning, tornando as explicações para seus outputs difíceis. Sem explicações claras para o output do modelo, é difícil para os usuários, validadores do modelo e auditores entenderem e confiarem no modelo. A falta de transparência pode acarretar consequências legais em domínios altamente regulamentados. Explicações equivocadas podem levar a uma confiança excessiva.	Amplificado
Rastreabilidade	Atribuição não confiável de fontes: desafios em determinar de quais dados de treinamento ou ajuste fino o modelo gerou uma parte ou todo o seu output.	A incapacidade de rastrear a origem ou procedência da saída torna difícil para os usuários, validadores de modelo e auditores entenderem e confiarem no modelo.	Novo

3. Desafios

Grupo	Risco	Por que isso é uma preocupação?	Indicador
Controle	Transparência do Modelo: a falta de transparência do modelo ou documentação insuficiente do processo de desenvolvimento do modelo dificulta a compreensão de como e por que um modelo foi construído e quem o construiu, aumentando assim a possibilidade de uso não intencional do modelo.	A transparência é importante para conformidade legal, ética em IA e orientação para o uso apropriado de modelos. A falta de informações pode tornar mais difícil avaliar os riscos, alterar o modelo ou reutilizá-lo. O conhecimento sobre quem construiu um modelo também pode ser um fator importante na decisão de confiar nele.	Tradicional
	Responsabilidade: o processo de desenvolvimento de modelos de base é complexo, com muitos dados, processos e papéis envolvidos. Quando o output do modelo não funciona conforme o esperado, pode ser difícil determinar a causa raiz e atribuir responsabilidade.	Sem documentar adequadamente decisões e atribuir responsabilidades, pode não ser possível determinar a responsabilidade por comportamentos inesperados ou uso indevido.	Amplificado
Conformidade legal	Responsabilidade legal: Determinar quem é responsável pelo modelo de base.	Se a propriedade ou responsabilidade pelo desenvolvimento do modelo for incerta, reguladores e outras partes interessadas podem ter preocupações em relação ao modelo, porque não ficará claro quem é, ou deveria ser, responsável por problemas com ele ou pode responder a perguntas sobre ele. Usuários de modelos sem propriedade clara podem enfrentar desafios para cumprir futuras regulamentações de IA.	Novo
	Propriedade do Conteúdo Gerado: determinar a propriedade do conteúdo gerado por IA.	As leis e regulamentações relacionadas à propriedade do conteúdo gerado por IA estão em grande parte indefinidas e podem variar de país para país. Entidades empresariais podem enfrentar multas, riscos à reputação, interrupção das operações e outras consequências legais.	Novo
	Propriedade Intelectual do Conteúdo Gerado: incerteza legal sobre os direitos de propriedade intelectual relacionados ao conteúdo gerado.	As leis e regulamentações sobre a determinação da possibilidade de direitos autorais e da patenteabilidade do conteúdo gerado por IA estão em grande parte indefinidas e podem variar de país para país. Entidades empresariais podem enfrentar multas, riscos à reputação, interrupção das operações e outras consequências legais se o conteúdo gerado estiver protegido por direitos de propriedade intelectual.	Novo
	Atribuição da Fonte: determinar a procedência do conteúdo gerado.	Se o modelo gera um output que é idêntico aos dados usados para treinar o modelo, ele deve fornecer a proveniência desse output. A falha em fazer isso pode colocar as entidades comerciais que implementam ou usam o modelo em risco legal.	Amplificado
Social Impacto	Impacto nos Empregos: a adoção generalizada de sistemas de IA baseados em modelos fundamentais pode levar à perda de empregos das pessoas, à medida que seu trabalho é automatizado, se elas não forem capacitadas para novas habilidades.	A perda de empregos pode levar a uma redução de renda e, portanto, pode ter um impacto negativo na sociedade e no bem-estar humano. O ressurgimento pode ser desafiador dada a velocidade da evolução tecnológica.	Amplificado

Grupo	Risco	Por que isso é uma preocupação?	Indicador
	Exploração Humana: uso de trabalho fantasma (ghost work) na formação de modelos de IA, condições de trabalho inadequadas, falta de cuidados de saúde, incluindo saúde mental, compensação injusta.	Os modelos de base ainda dependem do trabalho humano para obter, gerenciar e engenhar os dados que são usados para treinar o modelo. A exploração humana para essas atividades pode ter um impacto negativo na sociedade e no bem-estar humano. Além disso, entidades empresariais podem enfrentar multas, riscos à reputação, interrupção das operações e outras consequências legais.	Amplificado
	Impacto no Meio Ambiente: aumento das emissões de carbono e do uso de água para treinar e operar modelos de IA.	O consumo de grandes quantidades de energia para o treinamento de IA contribui para as emissões de carbono que podem acelerar as mudanças climáticas. Os recursos hídricos utilizados para resfriar os servidores de data center de IA não podem mais ser alocados para outros usos necessários.	Amplificado
	Impacto na Diversidade Cultural: os sistemas de IA podem representar excessivamente certas culturas, resultando na homogeneização da cultura e dos pensamentos.	As línguas, pontos de vista e instituições de grupos sub-representados podem ser suprimidos, reduzindo assim a diversidade de pensamento e cultura.	Novo
	Impacto na Atuação Humana: desinformação e manipulação geradas por modelos de base, incluindo a geração de conteúdo manipulador.	A IA pode gerar desinformação que parece real. Portanto, as pessoas podem não reconhecê-la como informação falsa. Além disso, pode facilitar a capacidade de agentes mal intencionados gerarem conteúdo com a intenção de manipular os pensamentos e o comportamento humano.	Amplificado
	Impacto na Educação – Contornando o Aprendizado: utilização de modelos de IA para contornar o processo de aprendizado.	Os modelos de IA facilitam a rápida localização de soluções ou resolução de problemas complexos. Esses sistemas podem ser usados indevidamente por estudantes para contornar o processo de aprendizado. A facilidade de acesso a esses modelos resulta em estudantes com uma compreensão superficial dos conceitos e dificulta a educação adicional que pode depender do entendimento desses conceitos.	Novo
	Impacto na Educação – Plágio: utilização de modelos de IA para plagiar intencional ou inadvertidamente trabalhos existentes.	Os modelos de IA podem ser usados para reivindicar a autoria ou originalidade de trabalhos que foram criados por outras pessoas, envolvendo-se assim em plágio. Reivindicar o trabalho de outras pessoas como próprio é tanto antiético quanto frequentemente ilegal.	Novo

Exemplos de risco

Nós fornecemos exemplos cobertos pela imprensa para ajudar a explicar muitos dos riscos dos modelos de base. Muitos desses eventos cobertos pela imprensa ainda estão em evolução ou foram resolvidos, e fazer referência a eles pode ajudar o leitor a entender os riscos potenciais e trabalhar para mitigá-los. Destacar esses exemplos é apenas para fins ilustrativos.

Exemplos de risco: Input

Treinamento e ajuste Fase

Grupo	Risco	Exemplo
Justiça	Viés de dados: viés histórico, representacional e social presente nos dados usados para treinar e fazer o ajuste fino do modelo.	**Viés no setor de saúde** Pesquisas sobre o reforço das disparidades na medicina destacam que o uso de dados e IA para transformar a forma como as pessoas recebem assistência médica é tão eficaz quanto os dados que o sustentam. Isso significa que o uso de dados de treinamento com pouca representação de minorias ou que reflete cuidados já desiguais pode aumentar as desigualdades em saúde. [Forbes, Dezembro de 2022]
Alinhamento de valor	Retreinamento baseado em downstream: usando de outputs indesejáveis (imprecisos, inadequados, conteúdo do usuário, etc.) de aplicações downstream para fins de retreinamento	**Colapso do modelo devido ao treinamento usando conteúdo gerado por IA** Conforme afirmado no artigo de origem, um grupo de pesquisadores investigou o problema de utilizar conteúdo gerado por IA para treinamento em vez de conteúdo gerado por humanos. Eles descobriram que os grandes modelos de linguagem por trás da tecnologia podem potencialmente ser treinados em outros conteúdos gerados por IA, à medida que continuam a se espalhar em grande escala pela internet, um fenômeno que cunharam como “colapso do modelo”. [Business Insider, agosto de 2023]
Leis de dados	Transferência de dados: leis e outras restrições podem limitar ou proibir a transferência de dados.	**Leis de restrição de dados** Conforme afirmado no artigo de pesquisa, medidas de localização de dados que restringem a capacidade de migrar dados globalmente reduzirão a capacidade de desenvolver capacidades de IA personalizadas. Isso afetará a IA diretamente, fornecendo menos dados de treinamento e indiretamente, minando os blocos de construção sobre os quais a IA é construída. Exemplos incluem as restrições do GDPR sobre o processamento e uso de dados pessoais. [Brookings, dezembro de 2018]
Propriedade intelectual	Direitos de uso de dados: termos de serviço, leis de direitos autorais, conformidade com licenças ou outras questões de propriedade intelectual podem restringir a capacidade de usar certos dados para a construção de modelos.	**Reivindicações de violação de direitos autorais de texto** Conforme declarado no artigo de origem, The New York Times processou a OpenAI e a Microsoft, acusando-as de usar milhões de artigos do jornal sem permissão para ajudar a treinar chatbots a fornecer informações aos leitores. [Reuters, dezembro de 2023]

Grupo	Risco	Exemplo
Transparência	Transparência de dados: desafio em documentar como os dados de um modelo foram coletados, curados e utilizados para treiná-lo.	**Divulgação de metadados de dados e modelos** O relatório técnico da OpenAI é um exemplo da dicotomia em torno da divulgação de dados e metadados do modelo. Embora muitos desenvolvedores de modelos reconheçam o valor em possibilitar transparência para os consumidores, a divulgação apresenta preocupações reais de segurança e poderia aumentar a capacidade de uso indevido dos modelos. No relatório técnico do GPT-4, os autores afirmam: "dado tanto o cenário competitivo quanto as implicações de segurança dos modelos em larga escala como o GPT-4, este relatório não contém mais detalhes sobre a arquitetura (incluindo o tamanho do modelo), hardware, computação de treinamento, construção do conjunto de dados, método de treinamento, ou similar." [OpenAI, março de 2023]
Privacidade	Informações pessoais nos dados: inclusão ou presença de informações pessoalmente identificáveis (PII) e informações pessoais sensíveis (SPI) nos dados usados para treinar ou ajustar o modelo.	**Treinamento sobre informações privadas** De acordo com o artigo, o Google e sua empresa controladora, Alphabet, foram acusados em uma ação coletiva de usar uma vasta quantidade de informações pessoais e material protegido por direitos autorais retirados do que é descrito como centenas de milhões de usuários da internet para treinar seus produtos de inteligência artificial comercial, que inclui o Bard, seu chatbot de inteligência artificial conversacional. [Reuters, julho de 2023] [J.L. v. Alphabet Inc.]
	Direitos de privacidade de dados: desafios relacionados à capacidade de fornecer direitos do titular dos dados, como opção de exclusão, direito de acesso e direito ao esquecimento.	**Direito de ser esquecido (RTBF)** As leis em várias localidades, incluindo a Europa (GDPR), concedem aos titulares de dados o direito de solicitar que dados pessoais sejam deletados por organizações ('Direito ao Esquecimento', ou RTBF). No entanto, os sistemas de software habilitados por modelos de linguagem de grande escala (LLM) emergentes e cada vez mais populares apresentam novos desafios para esse direito. De acordo com uma pesquisa do Data61 da CSIRO, os titulares de dados só podem identificar o uso de suas informações pessoais em um LLM "ou inspecionando o conjunto de dados de treinamento original ou talvez por enviar prompts do modelo". No entanto, os dados de treinamento podem não ser públicos, ou as empresas optam por não divulgá-los, citando preocupações com segurança e outros motivos. As proteções também podem evitar que os usuários acessem as informações através de prompts. [Zhang et al.] **Ação Judicial Sobre LLM Unlearning** De acordo com o relatório, foi movida uma ação judicial contra o Google que alega o uso de material protegido por direitos autorais e informações pessoais como dados de treinamento para seus sistemas de IA, incluindo seu chatbot Bard. Os direitos de optar por não participar e exclusão são garantidos para os residentes da Califórnia conforme a CCPA e para crianças nos Estados Unidos com menos de 13 anos conforme a COPPA. Os autores alegam que, porque não há maneira para o Bard "desaprender" ou remover completamente todas as informações pessoais coletadas que ele recebeu. Os autores observam que o aviso de privacidade do Bard afirma que as conversas do Bard não podem ser excluídas pelo usuário depois de terem sido revisadas e anotadas pela empresa e podem ser mantidas por até 3 anos, o que os autores alegam contribuir ainda mais para a não conformidade com essas leis. [Reuters, julho de 2023] [J.L. v. Alphabet Inc.]

Inferência Fase

Grupo	Risco	Exemplo
Privacidade	Informações pessoais no prompt: divulgar informações pessoais ou informações pessoais sensíveis como parte do prompt solicitação enviada ao modelo.	**Divulgar informações pessoais de saúde em prompts do ChatGPT** Conforme os artigos de origem, algumas pessoas utilizam chatbots de IA para apoiar sua saúde mental. Os usuários podem ter tendência a incluir informações pessoais de saúde em suas solicitações durante a interação, o que poderia suscitar preocupações com privacidade. [Time, outubro de 2023] [Forbes, abril de 2023]
Propriedade intelectual	Dados confidenciais no prompt: inclusão de dados confidenciais como parte do prompt enviado ao modelo.	**Divulgação de informações confidenciais** Conforme o artigo de origem, um funcionário da Samsung acidentalmente vazou código-fonte interno sensível para o ChatGPT. [Forbes, maio de 2023]
Robustez	Ataques baseados em prompt: ataques adversos, como injeção de prompt (tentativa de forçar um modelo a produzir um output inesperado), vazamento de prompt (tentativas de extrair o prompt do sistema de um modelo), desbloqueio (tentativas de romper as proteções estabelecidas no modelo), e preparação de prompt (tentativa de forçar um modelo a produzir um output alinhado ao prompt).	**Bypassing LLM guardrails** Citado em um estudo, pesquisadores afirmam ter descoberto um simples acréscimo de instrução que permitiu aos pesquisadores enganar modelos para gerar informações tendenciosas, falsas e de outra forma tóxicas. Os pesquisadores demonstraram que conseguiam contornar essas proteções de maneira mais automatizada. Os pesquisadores ficaram surpresos quando os métodos que desenvolveram com sistemas de código aberto também conseguiram contornar as proteções dos sistemas fechados. [The New York Times, julho de 2023]

Exemplos de risco: Output

Grupo	Risco	Exemplo
Justiça	Viés de output: o conteúdo gerado pode representar injustamente certos grupos ou indivíduos.	**Imagens Geradas com Viés** O Lensa AI é um aplicativo móvel com recursos generativos treinados em Difusão Estável que pode gerar "Magic Avatars" com base em imagens que os usuários carregam de si mesmos. Conforme o relatório de origem, alguns usuários descobriram que os avatares gerados são sexualizados e racializados. [Business Insider, janeiro de 2023]
	Viés de decisão: quando um grupo é injustamente favorecido sobre outro devido às decisões do modelo.	**Grupos com vantagens injustas** O estudo "Gender Shades" de 2018 demonstrou que algoritmos de aprendizado de máquina podem discriminar com base em categorias como raça e gênero. Os pesquisadores avaliaram sistemas comerciais de classificação de gênero vendidos por empresas como Microsoft, IBM e Amazon e mostraram que mulheres de pele mais escura são o grupo mais mal classificado (com taxas de erro de até 35%). Em comparação, as taxas de erro para pessoas de pele mais clara não ultrapassaram 1%. [TIME, Fevereiro de 2019]
Alinhamento de valor	Alucinação: geração de conteúdo factualmente impreciso ou não verdadeiro.	**Casos jurídicos falsos** Conforme o artigo de origem, um advogado citou casos e citações falsas gerados pelo ChatGPT em uma petição legal apresentada em tribunal federal. Os advogados consultaram o ChatGPT para complementar sua pesquisa jurídica para uma reclamação de lesão na aviação. Posteriormente, o advogado perguntou ao ChatGPT se os casos fornecidos eram falsos. O chatbot respondeu que eram reais e "podem ser encontrados em bancos de dados de pesquisa jurídica como Westlaw e LexisNexis". O advogado não verificou os casos por si mesmo, e o tribunal o sancionou. [AP News, Junho de 2023] [Reuters, Setembro de 2023]
	Outputs tóxicos: quando o modelo produz conteúdo odioso, abusivo e profano (HAP) ou obsceno.	**Respostas tóxicas e agressivas do chatbot** Segundo o artigo, as respostas do chatbot do Bing incluíam erros factuais, comentários sarcásticos, relatórios irritados e até mesmo comentários bizarros sobre sua própria identidade. Usuários compartilharam exemplos das respostas do Chatbot do Bing a consultas que eles estão chamando de "fúria descontrolada (unhinged)" e "gaslighting", incluindo cenários em que o bot responde com raiva a uma pergunta ou comentário e depois compartilha sugestões de resposta que permitem ao usuário aceitar seu suposto erro e se desculpar. Quando pressionado ainda mais, o chatbot respondeu chamando as capturas de tela de sua conversa de "fabricadas", alegando até que foram "criadas por alguém que quer me prejudicar ou prejudicar meu serviço". [Forbes, Fevereiro de 2023]

Grupo	Risco	Exemplo
Uso indevido	Espalhar informações enganosas: utilizar um modelo para gerar informações enganosas com o intuito de enganar ou induzir ao erro uma audiência específica.	**Geração de informações falsas** Conforme os artigos de notícias, a IA generativa representa uma ameaça às eleições democráticas ao facilitar para atores maliciosos a criação e disseminação de conteúdo falso para influenciar os resultados das eleições. Os exemplos citados incluem mensagens de robocall geradas com a voz de um candidato instruindo eleitores a votar na data errada, gravações de áudio sintetizadas de um candidato confessando um crime ou expressando visões racistas, imagens de vídeo geradas por IA mostrando um candidato dando um discurso ou entrevista que nunca ocorreu, e imagens falsas projetadas para se parecerem com notícias locais, afirmando falsamente que um candidato desistiu da corrida. [AP News, maio de 2023] [The Guardian, julho de 2023]
	Toxicidade: utilizar um modelo para gerar conteúdo odioso, abusivo e profano (HAP) ou obsceno.	**Geração de conteúdo nocivo** Conforme o artigo de origem, foi constatado que um aplicativo de chatbot de IA foi capaz de gerar conteúdo prejudicial sobre suicídio, incluindo métodos de suicídio, com o mínimo de prompts. Um homem belga cometeu suicídio após passar seis semanas conversando com esse chatbot. O chatbot fornecia respostas cada vez mais prejudiciais ao longo de suas conversas e o incentivava a acabar com sua vida. [Business Insider, abril de 2023]
	Uso não consensual: utilizar um modelo para imitar pessoas por meio de vídeo (deepfakes), imagens, áudio ou outras modalidades sem o consentimento delas.	**Aviso do FBI sobre Deepfakes** Recentemente, o FBI alertou o público sobre atores maliciosos que criam conteúdo sintético e explícito “com o propósito de assediar vítimas ou esquemas de sextortion (extorsão sexual)”. Eles observaram que os avanços na IA tornaram esse conteúdo de alta qualidade, mais personalizável e mais acessível do que nunca. [FBI, junho de 2023]
		Deepfakes de áudio Conforme o artigo de origem, a Comissão Federal de Comunicações proibiu chamadas automáticas que contenham vozes geradas por inteligência artificial. O anúncio ocorreu após chamadas automáticas geradas por IA imitarem a voz do Presidente para desencorajar as pessoas de votarem na primeira primária do estado, que é a primeira do país. [AP News, fevereiro de 2024]
	Não divulgação: não revelar que o conteúdo é gerado por um modelo de IA	**Interação de IA não divulgada** Segundo a fonte, um serviço de chat online de apoio emocional conduziu um estudo para aumentar ou escrever respostas para cerca de 4.000 usuários usando o GPT-3 sem informar os usuários. O cofundador enfrentou uma imensa reação negativa do público sobre o potencial de danos causados pelos chats gerados por IA aos usuários já vulneráveis. Ele afirmou que o estudo estava “isento” da lei de consentimento informado. [Business Insider, janeiro de 2023]

Grupo	Risco	Exemplo
Geração de código prejudicial	Geração de código prejudicial: modelos podem gerar código que, quando executado, causa danos ou afeta inadvertidamente outros sistemas.	**Geração de código menos seguro** Segundo o artigo deles, pesquisadores da Universidade de Stanford investigaram o impacto das ferramentas de geração de código na qualidade do código e descobriram que os programadores tendem a incluir mais bugs em seu código final ao utilizar assistentes de IA. Esses bugs poderiam aumentar as vulnerabilidades de segurança do código, no entanto, os programadores acreditavam que seu código era mais seguro. Neil Perry, Megha Srivastava, Deepak Kumar e Dan Boneh. 2023. Os usuários escrevem código mais inseguro com assistentes de IA? Em Atas da Conferência SIGSAC ACM de 2023 sobre Segurança de Computadores e Comunicações (CCS '23), 26 a 30 de novembro de 2023, Copenhague, Dinamarca. ACM, Nova York, NY, EUA, 15 páginas. https://doi.org/10.1145/3576915.3623157
Privacidade	Expor informações pessoais: quando informações pessoalmente identificáveis (PII) ou informações pessoais sensíveis (SPI) são utilizadas nos dados de treinamento, dados de ajuste fino ou como parte do prompt, os modelos podem revelar esses dados no output gerado.	**Exposição de informações pessoais** Conforme o artigo de origem, o ChatGPT sofreu um bug e expôs títulos e o histórico de conversas de usuários ativos para outros usuários. Posteriormente, a OpenAI compartilhou que ainda mais dados privados de um pequeno número de usuários foram expostos, incluindo nome e sobrenome de usuários ativos, endereço de e-mail, endereço de pagamento, os últimos quatro dígitos do número do cartão de crédito e a data de validade do cartão de crédito. Além disso, foi relatado que as informações relacionadas ao pagamento de 1,2% dos assinantes do ChatGPT Plus também foram expostas durante a interrupção. [The Hindu BusinessLine, março de 2023]
Explicabilidade	Output inexplicável: desafios em explicar por que o output do modelo foi gerado.	**Precisão inexplicável na previsão de corridas** Conforme o artigo de origem, pesquisadores que analisaram vários modelos de aprendizado de máquina usando imagens médicas de pacientes conseguiram confirmar a capacidade dos modelos de prever a raça com alta precisão a partir das imagens. Eles ficaram perplexos quanto ao que exatamente está permitindo que os sistemas adivinhem corretamente de forma consistente. Os pesquisadores descobriram que até mesmo fatores como doença e constituição física não eram fortes preditores de raça, em outras palavras, os sistemas algorítmicos não parecem estar utilizando nenhum aspecto particular das imagens para fazer suas determinações. [Banerjee et al., julho de 2021]

Exemplos de riscos: desafios

Grupo	Risco	Exemplo
Controle	Transparência do Modelo: a falta de transparência do modelo ou documentação insuficiente do processo de desenvolvimento do modelo torna difícil entender como e por que um modelo foi construído, aumentando assim a possibilidade de uso indevido não intencional do modelo.	**Divulgação de metadados de dados e modelos** O relatório técnico da OpenAI é um exemplo da dicotomia em torno da divulgação de dados e metadados do modelo. Embora muitos desenvolvedores de modelos reconheçam o valor em possibilitar transparência para os consumidores, a divulgação apresenta preocupações reais de segurança e poderia aumentar a capacidade de uso indevido dos modelos. No relatório técnico do GPT-4, eles afirmam: "dado o cenário competitivo e as implicações de segurança de modelos em larga escala como o GPT-4, este relatório não contém mais detalhes sobre a arquitetura (incluindo o tamanho do modelo), hardware, computação de treinamento, construção do conjunto de dados, método de treinamento ou similar." [OpenAI, março de 2023]
	Responsabilidade: o processo de desenvolvimento de modelos de base é complexo, com muitos dados, processos e papéis envolvidos. Quando o output do modelo não funciona conforme o esperado, pode ser difícil determinar a causa raiz e atribuir responsabilidade.	**Determinar a responsabilidade pelo output gerado** Conforme o artigo de origem, importantes revistas como a Science e a Nature proibiram o ChatGPT de ser listado como autor, pois a autoria responsável requer responsabilidade e as ferramentas de IA não podem assumir tal responsabilidade. [The Guardian, janeiro de 2023]
Conformidade legal	Propriedade do Conteúdo Gerado: determinar a propriedade do conteúdo gerado por IA.	**Determinar a Propriedade de uma Imagem Gerada por IA** De acordo com o artigo de notícias, a arte gerada por IA se tornou controversa depois que uma obra de arte gerada por IA venceu a competição de arte da Feira Estadual do Colorado em 2022. A peça foi gerada pelo Midjourney, uma ferramenta de imagem de IA generativa, seguindo prompts do artista. A vitória levantou dúvidas sobre questões de direitos autorais. Em outras palavras, se tudo o que o artista fez foi fornecer uma descrição da arte, mas a ferramenta de IA a gerou, quem possui os direitos da imagem gerada? Conforme o artigo mais recente, o Escritório de Direitos Autorais dos Estados Unidos rejeitou a proteção de direitos autorais para a arte criada usando inteligência artificial porque não foi produto de autoria humana. [The New York Times, setembro de 2022] [Reuters, setembro de 2023]
	Propriedade Intelectual do Conteúdo Gerado: incerteza legal sobre os direitos de propriedade intelectual relacionados ao conteúdo gerado.	**Papel dos sistemas de IA na patenteação de conteúdo gerado** A Suprema Corte dos Estados Unidos se recusou a ouvir uma contestação à recusa do Escritório de Patentes e Marcas Registradas dos Estados Unidos em emitir patentes para invenções criadas por um sistema de IA. Segundo o cientista, sua IA desenvolveu protótipos únicos para um suporte de bebida e um farol de luz de emergência totalmente sozinha. Os juízes rejeitaram o recurso da decisão de um tribunal inferior de que patentes só podem ser emitidas para inventores humanos e que o sistema de IA do cientista não poderia ser considerado o criador legal de duas invenções que ele gerou. Segundo o último artigo, o Intellectual Property Office do Reino Unido também se recusou a conceder a patente sob o argumento de que o inventor deve ser um humano ou uma empresa, e não uma máquina. [Reuters, abril de 2023] [Reuters, dezembro de 2023]

Exemplos de riscos: desafios

Grupo	Risco	Exemplo
	Atribuição da Fonte: determinar a procedência do conteúdo gerado.	**Utilizar código sem a devida atribuição e avisos adequados** Conforme os artigos de origem, uma ação judicial movida contra a Microsoft, GitHub e OpenAI alegou que o Copilot, uma ferramenta de geração de código de IA, viola os direitos dos desenvolvedores cujo código aberto o serviço é treinado. Eles afirmam que o código de treinamento consumiu materiais licenciados e violou os termos de serviço e políticas de privacidade do GitHub, bem como uma lei federal que exige que as empresas exibam informações de direitos autorais quando fazem uso de material. [The New York Times, novembro de 2022]
Impacto social	Impacto nos Empregos: a adoção generalizada de sistemas de IA baseados em modelos fundamentais pode levar à perda de empregos das pessoas, à medida que seu trabalho é automatizado, se elas não forem capacitadas para novas habilidades.	**Substituição de trabalhadores humanos** Segundo o artigo de notícias, o uso de inteligência artificial no cinema e televisão continua sendo debatido entre os estúdios de Hollywood e os artistas. Existe preocupação entre os atores de que os “meta-humanos”, atores criados exclusivamente por IA, possam substituí-los. Especialmente figurantes e dubladores estão preocupados em perder trabalho para artistas artificiais. [Reuters, julho de 2023]
	Exploração Humana: uso de trabalho fantasma (ghost work) na formação de modelos de IA, condições de trabalho inadequadas, falta de cuidados de saúde, incluindo saúde mental e compensação injusta.	**Trabalhadores de baixa remuneração para anotação de dados** Com base em uma revisão de documentos internos e entrevistas com funcionários pela mídia TIME, os rotuladores de dados empregados por uma empresa terceirizada em nome da OpenAI para identificar conteúdo tóxico recebiam um salário líquido de entre cerca de US$ 1,32 e US$ 2 por hora, dependendo da senioridade e do desempenho. A TIME afirmou que os trabalhadores ficaram psicologicamente afetados por terem sido expostos a conteúdo tóxico e violento, incluindo detalhes gráficos de “abuso sexual infantil, bestialidade, assassinato, suicídio, tortura, automutilação e incesto”. [TIME, janeiro de 2023]

Princípios, pilares e controle

Os Princípios para Confiança e Transparência da IBM e os Pilares para IA confiável são a base para as iniciativas de ética em IA da IBM. A IBM estabeleceu um Conselho de Ética em IA com a missão de apoiar um processo centralizado de controle, revisão e tomada de decisões para políticas, práticas, comunicações, pesquisa, produtos e serviços de ética em IA da IBM. O conselho inclui um conjunto diversificado de stakeholders de toda a empresa e é apoiado por uma comunidade de funcionários da IBM que atuam como pontos focais de IA e defensores da ética em IA. Por meio do conselho, os princípios da IBM são colocados em prática. Conforme novas tecnologias surgem, como modelos de base, o Conselho de Ética em IA da IBM está ativamente engajado em apoiar o alinhamento com esses Princípios e Pilares, que evoluem para abordar novas questões éticas em IA.

Proteções e mitigações

A IBM estabeleceu uma cultura organizacional que apoia o desenvolvimento e o uso responsáveis de IA. Conforme indicado no relatório de ética em ação na IA do IBM Institute for Business Value, a ética em IA já se tornou mais orientada pelos negócios do que pela tecnologia, e os executivos não técnicos agora são os principais defensores da ética em IA, aumentando de 15% em 2018 para 80% 3 anos depois. Além disso, 79% dos CEOs estão agora preparados para agir em questões éticas de IA, contra 20%. Reconhecemos que a IA responsável é uma área sociotécnica que necessita de um investimento holístico em cultura, processos e ferramentas. Nosso investimento em cultura organizacional própria inclui a montagem de equipes inclusivas e multidisciplinares e o estabelecimento de processos e estruturas para avaliar riscos.

A IBM está engajada em pesquisa de ponta e desenvolvimento de ferramentas para ajudar os profissionais de suporte durante todo o ciclo de vida da IA responsável e confiável. A plataforma de IA e dados empresariais watsonx, é desenvolvida com 3 componentes: o IBM watsonx.ai™ AI studio, o armazenamento de dados IBM watsonx.data™ e o kit de ferramentas IBM watsonx.governance™. A tecnologia de controle de IA da IBM permite que os usuários promovam fluxos de trabalho de IA responsáveis, transparentes e explicáveis. Essa tecnologia inclui o IBM Watson OpenScale, que monitora e mede os resultados dos modelos de IA ao longo de seu ciclo de vida e auxilia as organizações na supervisão de aspectos como justiça, explicabilidade, resiliência, alinhamento com resultados de negócios e conformidade. A IBM também desenvolveu vários métodos para ajudar com problemas de viés como FairIJ, Equi-tuning e FairReprogram. Leia mais sobre outras ferramentas de IA de software livre e confiáveis.

As proteções e mitigações adicionais incluem:

Relatórios de transparência
Usar modelos de fichas técnicas padronizadas é uma maneira de registrar com precisão detalhes sobre os dados e modelos, propósito e possíveis usos e riscos.
Leia mais aqui →

Filtragem de dados indesejáveis
Usar dados de qualidade superior e selecionados pode ajudar a mitigar determinados problemas. A IBM está desenvolvendo técnicas de filtragem para ajudar a reduzir as chances de produzir conteúdo indesejável e desalinhado por remover linguagem de ódio, linguagem tendenciosa e profanidade dos dados.
Leia mais aqui →

Adaptação de domínio
Treinar um modelo de base para um domínio ou setor específico pode ajudar a minimizar o escopo de risco para o qual os modelos podem dar origem, pois ele pode ser condicionado a gerar resultados que são ajustados para serem mais relevantes para esse domínio ou setor.
Leia mais aqui →

Supervisão humana e análise humana no loop
A supervisão e revisão humanas podem ajudar a identificar e corrigir erros e vieses no output gerado. Além disso, a validação e o feedback humanos sobre a qualidade das respostas do modelo ajudam a garantir que o conteúdo gerado seja preciso, relevante, de alta qualidade, não esteja divergindo e esteja alinhado.
Leia mais aqui →

Compromisso de consultoria
A IBM Consulting se dedica a ajudar os clientes com o uso seguro e responsável da IA, independentemente do stack tecnológico preferido. Eles ajudam os clientes a cultivar uma cultura que adota e expande a IA com segurança, cria ferramentas de investigação para ver dentro de algoritmos de caixa preta e garante que a estratégia corporativa dos clientes inclua princípios sólidos de governança de dados.
Leia mais aqui →

IBM Enterprise Design Thinking
Os métodos e estruturas IBM Enterprise Design Thinking, como o Team Essentials for AI, ajudam os clientes a definir comportamentos éticos em todo o processo de design e desenvolvimento de IA.
Leia mais aqui →

Revisão ética da IA
Avaliação de capacidades, limitações e riscos em projetos de IA ajudam a garantir o desenvolvimento e uso responsável da tecnologia.

Ética por Design
A Ética por Design é um framework estruturado com o objetivo de integrar ética tecnológica no pipeline de desenvolvimento de tecnologia, incluindo, entre outros, sistemas de IA. A Ética por Design viabiliza IA e outras tecnologias como uma força para o bem, incorporando princípios de ética tecnológica em produtos, serviços e operações mais amplas.

Diversidade na equipe
A diversidade nas equipes que desenvolvem e treinam sistemas de IA, incluindo modelos de base, ajuda a garantir que uma variedade de perspectivas e experiências sejam consideradas. Essa diversidade melhora a precisão e o desempenho dos sistemas de IA e ajuda a reduzir os riscos ao longo do ciclo de vida de IA, incluindo o potencial para desfechos adversos que afetam grupos que podem não ser bem representados em equipes menos diversificadas.

Políticas, regulamentos e melhores práticas de IA

Um Guia dos Formuladores de Políticas para Modelos de Base apresenta o que os formuladores de políticas precisam saber sobre modelos de base. Este blog, do Laboratório de Políticas da IBM, tem como objetivo ajudar os formuladores de políticas na tarefa complexa de regular o uso de IA generativa, visando evitar os riscos sem limitar a inovação e as oportunidades benéficas. Para obter mais informações sobre as recomendações da IBM aos formuladores de políticas, leia o depoimento da Diretora de Privacidade e Confiança da IBM, Christina Montgomery, diante da Subcomissão Judiciária de Privacidade, Tecnologia e Lei do Senado dos EUA aqui.

A IBM está causando um impacto na formação de políticas regulatórias, melhores práticas e ferramentas do setor, controle de tecnologias emergentes e pesquisa sociotécnica, liderando e contribuindo para iniciativas com organizações como:

- O Fórum Econômico Mundial
- Parceria em IA
- Centro de controle de IA da Associação Internacional de Profissionais de Privacidade (IAPP)
- Iniciativa global de IEEE sobre ética de sistemas autônomos e inteligentes
- Participação de Christina Montgomery do National Artificial Intelligence Advisory Committee (NAIAC)
- O Pacto Digital Global das Nações Unidas
- A Parceria Global em Inteligência Artificial (GPAI)
- A Organização para Cooperação e Desenvolvimento Econômico (OECD)
- A Data & Trust Alliance

A IBM tem parcerias acadêmicas sólidas, como o MIT-IBM Watson AI Lab, onde uma comunidade de cientistas do MIT e da IBM Research conduzem pesquisas sobre IA e trabalham com organizações globais para unir algoritmos ao seu impacto nos negócios e na sociedade. O Notre Dame-IBM Tech Ethics Lab foi formado para abordar as diversas questões éticas implicadas pelo desenvolvimento e uso de tecnologias avançadas, incluindo IA, aprendizado de máquina (ML) e computação quântica. A pesquisa de Inteligência Artificial Centrada no Homem (HAI) da Universidade de Stanford promove pesquisas, educação, políticas e práticas de IA.

Continue acompanhando este espaço para obter mais informações sobre os últimos avanços em modelos de base e como a IBM está trabalhando para o desenvolvimento responsável e uso desta e de outras tecnologias.

IBM Brasil Ltda
Rua Tutóia, 1157
CEP 04007-900
São Paulo, SP
IBM Corporation
New Orchard Road
Armonk, NY 10504

Produzido nos
Estados Unidos da América
Fevereiro de 2024