AI ACTION SUMMIT

Relatório internacional sobre a segurança da IA

Relatório científico internacional sobre a segurança da IA avançada

Janeiro de 2025

Colaboradores

PRESIDENTE

Prof. Yoshua Bengio, Universidade de Montreal/Mila - Instituto de IA de Quebec

PAINEL CONSULTIVO DE ESPECIALISTAS

Este painel internacional foi indicado pelos governos dos 30 países listados a seguir, pela ONU, pela UE e pela OCDE.

Alemanha: Antonio Krüger, Centro Alemão de Pesquisa em Inteligência Artificial

Austrália: Bronwyn Fox, Universidade de Nova Gales do Sul

Brasil: André Carlos Ponce de Leon Ferreira de Carvalho, Instituto de Matemática e Informática, Universidade de São Paulo

Canadá: Mona Nemer, Assessora científica principal do Canadá

Chile: Raquel Pezoa Rivera, Universidade Técnica Federico Santa María

China: Yi Zeng, Academia de Ciências da China

Emirados Árabes Unidos: Marwan Alserkal, Ministério de Assuntos do Gabinete, Gabinete do Primeiro-Ministro

Espanha: Nuria Oliver, ELLIS Alicante

Estados Unidos: El Saif M. Khan, Ministério do Comércio dos EUA

Filipinas: Dominic Vincent Ligot, CirroLytix

França: Guillaume Avrin, Coordenador Nacional de Inteligência Artificial

Índia: Balaraman Ravindran, Escola Wadhwani de Ciência de Dados e Inteligência Artificial, Instituto Indiano de Tecnologia de Madras

Indonésia: Hammam Riza, Pesquisa Colaborativa e Inovação Industrial em Inteligência Artificial (KORIKA)

Irlanda: Ciarán Seoighe, Research Ireland

Israel: Ziv Katzir, Autoridade de Inovação de Israel

Itália: Andrea Monti, Especialista jurídico da Subsecretaria de Estado para a Transformação Digital, Presidência do Conselho de Ministros Italiano

Japão: Hiroaki Kitano, Sony Group Corporation

Quênia: Nusu Mwamanzi, Ministério de TIC e Economia Digital

México: José Ramón López Portillo, LobsterTel

Nações Unidas: Amandeep Singh Gill, Subsecretário-Geral de Tecnologias Digitais e Emergentes e Enviado do Secretário-Geral para a Tecnologia

Nigéria: Olubunmi Ajala, Ministério das Comunicações, Inovação e Economia Digital

Nova Zelândia: Gill Jolly, Ministério de Negócios, Inovação e Emprego

OCDE: Jerry Sheehan, Diretor da Diretoria de Ciência, Tecnologia e Inovação

Países Baixos: Haroon Sheikh, Conselho Científico de Políticas Governamentais dos Países Baixos

República da Coreia: Kyoung Mu Lee, Departamento de Engenharia Elétrica e Informática, Universidade Nacional de Seul

Ruanda: Crystal Rugege, Centro para a Quarta Revolução Industrial

Singapura: Denise Wong, Grupo de Inovação e Proteção de Dados, Autoridade de Desenvolvimento de Mídia de Infocomm

Suíça: Christian Busch, Departamento Federal de Assuntos Econômicos, Educação e Pesquisa

Turquia: Ahmet Halit Hatip, Ministério da Indústria e Tecnologia da Turquia

Ucrânia: Oleksii Molchanovskyi, Comitê de Especialistas sobre o Desenvolvimento da Inteligência Artificial na Ucrânia

União Europeia: Juha Heikkilä, Escritório Europeu de IA

Reino da Arábia Saudita: Fahad Albalawi, Autoridade Saudita de Dados e Inteligência Artificial

Reino Unido: Chris Johnson, Assessor científico chefe do Departamento de Ciência, Inovação e Tecnologia

DIRETOR CIENTÍFICO

Sören Mindermann, Mila – Instituto de IA de Quebec

REDATOR PRINCIPAL

Daniel Privitera, Centro KIRA

GRUPO DE REDAÇÃO

Tamay Besiroglu, Epoch AI

Rishi Bommasani, Universidade de Stanford

Stephen Casper, Instituto de Tecnologia de Massachusetts

Yejin Choi, Universidade de Stanford

Philip Fox, Centro KIRA

Ben Garfinkel, Universidade de Oxford

Danielle Goldfarb, Mila - Instituto de Inteligência Artificial de Quebec

Hoda Heidari, Universidade Carnegie Mellon

Anson Ho, Epoch AI

Sayash Kapoor, Universidade de Princeton

Leila Khalatbari, Universidade de Ciência e Tecnologia de Hong Kong

Shayne Longpre, Instituto de Tecnologia de Massachusetts

Sam Manning, Centro para a Governança da IA

Vasilios Mavroudis, Instituto Alan Turing

Mantas Mazeika, Universidade de Illinois em Urbana-Champaign

Julian Michael, Universidade de Nova York

Jessica Newman, Universidade da Califórnia, Berkeley

Kwan Yee Ng, Concordia AI

Chinasa T. Okolo, Instituição Brookings

Deborah Raji, Universidade da Califórnia, Berkeley

Girish Sastry, independente

ASSESSORES SÊNIORES

Daron Acemoglu, Instituto de Tecnologia de Massachusetts

Olubayo Adekanmbi, contribuiu como assessor sênior antes de assumir seu cargo na EqualyzAI

David Dalrymple, Agência de Pesquisa e Invenção Avançada

Tomás G. Dietterich, Universidade Estadual do Oregon

Eduardo W. Fieltro, Universidade de Princeton

Pascale Fung, contribuiu como assessora sênior antes de assumir seu cargo na Meta

Pierre-Olivier Gourinchas, Departamento de Pesquisa, Fundo Monetário Internacional

Fredrik Heintz, Universidade de Linköping

Geoffrey Hinton, Universidade de Toronto

Nick Jennings, Universidade de Loughborough

Andreas Krause, ETH Zurich

Susan Leavy, University College Dublin

Percy Liang, Universidade de Stanford

Teresa Ludermir, Universidade Federal de Pernambuco

Vidushi Marda, AI Collaborative

Elizabeth Seger (redatora generalista), Demos

Theodora Skeadas, Humane Intelligence

Tobin South, Instituto de Tecnologia de Massachusetts

Emma Strubell, Universidade Carnegie Mellon

Florian Tramèr, ETH Zurich

Lucía Velasco, Universidade de Maastricht

Nicole Wheeler, Universidade de Birmingham

Helen Margetts, Universidade de Oxford

John McDermid, Universidade de York

Jane Munga, Fundação Carnegie para a Paz Internacional

Arvind Narayanan, Universidade de Princeton

Alondra Nelson, Instituto de Estudos Avançados

Clara Neppel, IEEE

Alice Oh, Escola de Informática KAIST

Gopal Ramchurn, responsável por IA no Reino Unido

Stuart Russell, Universidade da Califórnia, Berkeley

Marietje Schaake, Universidade de Stanford

Bernhard Schölkopf, ELLIS Institute Tübingen

Dawn Song, Universidade da Califórnia, Berkeley

Alvaro Soto, Pontifícia Universidade Católica do Chile

Lee Tiedrich, Universidade de Duke

Gaël Varoquaux, Inria

Andrew Yao, Instituto de Ciências da Informação Interdisciplinar, Universidade de Tsinghua

Ya-Qin Zhang, Universidade de Tsing

SECRETARIA

Instituto de Segurança de IA

Baran Acar

Ben Clifford

Rian Overy

Ben Snodin

Mila - Instituto de Inteligência Artificial de Quebec

Lambrini Das

Claire Dennis

Freya Hempleman

Hannah Merchant

Jonathan Barry

Benjamin Prud'homme

AGRADECIMENTOS

Revisores da sociedade civil e da indústria

Sociedade civil: Instituto Ada Lovelace, AI Forum New Zealand/Te Kāhui Atamai Iahiko o Aotearoa, Grupo Temporário de Especialistas em IA da Austrália, Fundação Carnegie para a Paz Internacional, Centro de Direito e Inovação/Fundação Certa, Centro para a Governança da IA, Centro de Direito e Inovação Digital do Presidente da Suprema Corte Meir Shamgar, Eon Institute, Gradient Institute, Israel Democracy Institute, Fundação Mozilla, Old Ways New, RAND, SaferAI, O Centro para a Resiliência de Longo Prazo, The Future Society, The Alan Turing Institute, The Royal Society, Associação de Políticas de Inteligência Artificial da Turquia.

Indústria: Advai, Anthropic, Cohere, Deloitte Consulting USA e Deloitte LLM UK, G42, Google DeepMind, Harmony Intelligence, Hugging Face, IBM, Lelapa AI, Meta, Microsoft, Shutterstock, Zhipu.ai.

Agradecimentos especiais

A Secretaria agradece o apoio, os comentários e as contribuições de Angie Abdilla, Nitarshan Rajkumar, Geoffrey Irving, Shannon Vallor, Rebecca Finlay e Andrew Strait.

© Direitos autorais da Coroa 2025

Esta publicação está sujeita aos termos da versão 3.0 da Licença de Governo Aberto, salvo indicação em contrário. Para consultar esta licença, visite nationalarchives.gov.uk/doc/open-government-licence/version/3 ou escreva para a Equipe de Política de Informação, The National Archives, Kew, London TW9 4DU, ou envie um e-mail para: psi@nationalarchives.gsi.gov.uk.

Quando identificarmos informações protegidas por direitos autorais de terceiros, você deverá obter permissão dos respectivos titulares desses direitos.

Qualquer consulta sobre esta publicação deve ser enviada para: secretariat.AIStateofScience@dsit.gov.uk .

As consultas sobre o conteúdo do relatório também devem ser enviadas ao responsável científico . .

Aviso de isenção de responsabilidade

O relatório não representa as opiniões do Presidente, de qualquer pessoa dos grupos de redação ou de assessoria, nem de qualquer um dos governos que apoiaram sua elaboração. Este relatório é uma síntese da pesquisa existente sobre as capacidades e os riscos da IA avançada. O Presidente do relatório é seu responsável máximo e supervisionou seu desenvolvimento do início ao fim.

Número de série da pesquisa: DSIT 2025/001

Sobre este relatório

● **Este é o primeiro Relatório internacional sobre a segurança da IA. Após uma publicação** provisória em maio de 2024, um grupo diversificado de 96 especialistas em inteligência artificial (IA) contribuiu para este primeiro relatório completo, incluindo um Painel consultivo de especialistas internacionais indicado por 30 países, pela Organização para a Cooperação e Desenvolvimento Econômico (OCDE), pela União Europeia (UE) e pelas Nações Unidas (ONU). O relatório pretende fornecer informações científicas que apoiem a formulação de políticas bem fundamentadas. Não recomenda políticas específicas.

● O relatório é obra de especialistas independentes. Os especialistas independentes que redigiram coletivamente este relatório, liderados pelo Presidente, tiveram plena autonomia sobre seu conteúdo.

● **Embora este relatório se concentre nos riscos e na segurança da IA, a IA também oferece muitos benefícios potenciais para as pessoas, as empresas e a sociedade. Existem** muitos tipos de IA, cada um com diferentes benefícios e riscos. Na maioria das vezes, e na maioria das aplicações, a IA aumenta a eficácia das pessoas e das organizações. Mas o mundo só poderá desfrutar plenamente e com segurança dos numerosos benefícios potenciais da IA se seus riscos forem gerenciados adequadamente. Este relatório se concentra em identificar esses riscos e avaliar métodos para mitigá-los. Não tem como finalidade avaliar exaustivamente todos os possíveis impactos sociais da IA, nem seus numerosos benefícios potenciais.

● O relatório se concentra na IA de propósito geral. O relatório se limita a um tipo de IA que avançou de forma especialmente rápida nos últimos anos e cujos riscos associados foram menos estudados e compreendidos: a IA de propósito geral, ou a IA que pode realizar uma ampla variedade de tarefas. A análise deste relatório se concentra nos sistemas de IA de propósito geral mais avançados no momento de sua redação, assim como em sistemas futuros que poderiam ser ainda mais capazes.

● **O relatório resume as evidências científicas em três perguntas fundamentais: O que pode** fazer a IA de propósito geral? Quais são os riscos associados à IA de propósito geral? E quais técnicas de mitigação existem diante desses riscos?

● Há muito em jogo. Nós, os especialistas que contribuímos para este relatório, continuamos a discordar sobre diversas questões, menores e maiores, relativas às capacidades da IA de propósito geral, aos riscos e à mitigação de riscos. No entanto, consideramos que este relatório é essencial para melhorar nossa compreensão coletiva dessa tecnologia e de seus riscos potenciais. Esperamos que o relatório ajude a comunidade internacional a avançar rumo a um maior consenso sobre a IA de propósito geral e a mitigar seus riscos de maneira mais eficaz, para que as pessoas possam experimentar seus numerosos benefícios potenciais de forma segura. Há muito em jogo. Esperamos dar continuidade a esta iniciativa.

Atualização sobre os últimos avanços em IA após a redação deste relatório: Nota do Presidente

Desde o fim da redação deste relatório (5 de dezembro de 2024) até sua publicação em janeiro de 2025, ocorreu um acontecimento importante. A empresa de inteligência artificial OpenAI compartilhou os primeiros resultados dos testes de um novo modelo de IA, o3. Esses resultados indicam um desempenho consideravelmente superior ao de qualquer modelo anterior em vários dos testes mais desafiadores das áreas de programação, raciocínio abstrato e raciocínio científico. Em alguns desses testes, o3 supera muitos especialistas humanos (mas não todos). Além disso, alcança um grande avanço em um teste fundamental de raciocínio abstrato que muitos especialistas, inclusive eu, acreditávamos estar fora do alcance dos sistemas de IA até pouco tempo atrás. No entanto, no momento da redação deste artigo, não há informações públicas sobre suas capacidades no mundo real, em particular para resolver tarefas mais abertas.

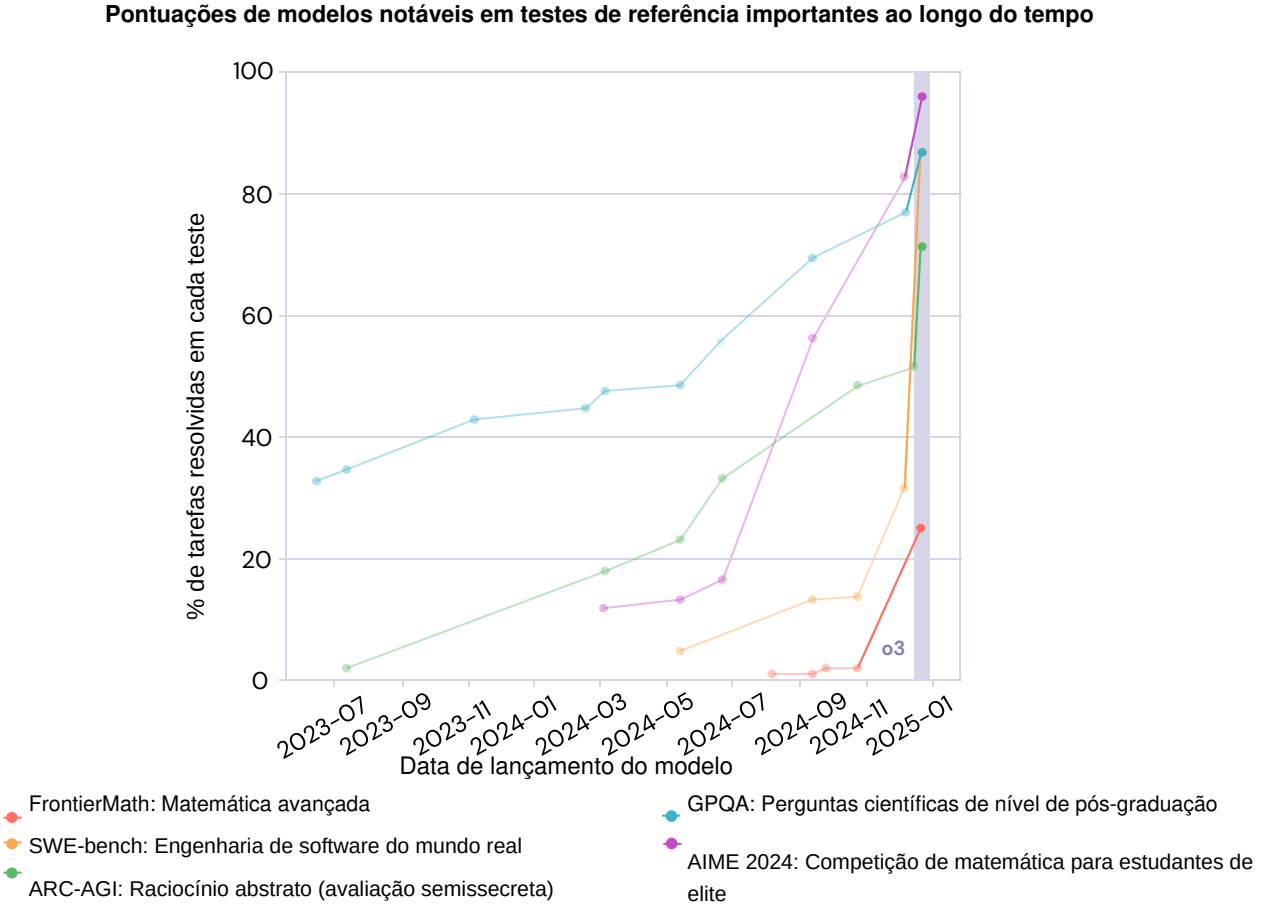


***Figura 0.1**: Pontuações de modelos de IA de propósito geral notáveis em testes de referência importantes de junho de 2023 a dezembro de 2024. o3 demonstrou uma melhora significativa de seu desempenho em comparação com o anterior (área sombreada). Esses testes de referência estão entre os mais desafiadores das áreas de programação, raciocínio abstrato e raciocínio científico. Para o modelo o3, que não foi lançado, é mostrada a data do anúncio; para os outros modelos, é mostrada a data de lançamento. Alguns dos modelos de IA mais recentes, incluindo o3, se*

beneficiaram de uma estrutura de apoio aprimorada e de mais computação durante o teste. Fontes: Anthropic, 2024; Chollet, 2024; Chollet et al., 2025; Epoch AI, 2024; Glazer et al. 2024; OpenAI, 2024a; OpenAI, 2024b; Jimenez et al., 2024; Jimenez et al., 2025.

Os resultados de o3 evidenciam que o ritmo dos avanços nas capacidades da IA pode continuar alto ou até mesmo se acelerar. Mais especificamente, sugerem que aumentar a capacidade computacional dos modelos para resolver um problema determinado ("escalonamento de inferência") ajuda a superar as limitações anteriores. Em termos gerais, o escalonamento de inferência aumenta o custo de uso dos modelos. No entanto, como demonstra outro modelo notável lançado recentemente, o R1, da empresa DeepSeek, em janeiro de 2025, os pesquisadores estão conseguindo reduzir esses custos. De modo geral, o escalonamento de inferência poderia permitir que os desenvolvedores de IA alcançassem mais avanços no futuro. Os resultados de o3 também destacam a necessidade de compreender melhor como o uso crescente da IA pelos desenvolvedores de IA pode afetar a velocidade do desenvolvimento futuro da própria IA.

As tendências evidenciadas por o3 poderiam ter implicações importantes para os riscos da IA. Os avanços anteriores na ciência e nas capacidades de programação geraram mais evidências sobre riscos como ataques cibernéticos e biológicos. Os resultados de o3 também são relevantes para os possíveis impactos no mercado de trabalho, o risco de perda de controle e o uso de energia, entre outros. Mas as capacidades de o3 também poderiam ser utilizadas para ajudar a fornecer proteção contra falhas de funcionamento e usos maliciosos. De modo geral, as avaliações de risco deste relatório devem ser analisadas levando em conta que a IA adquiriu novas capacidades desde que o relatório foi redigido. No entanto, até o momento, ainda não há evidências sobre os impactos reais de o3, nem informações para confirmar ou descartar novos riscos importantes ou riscos importantes imediatos.

A melhora das capacidades sugerida pelos resultados de o3 e nossa compreensão limitada das implicações para os riscos da IA ressaltam um desafio fundamental para os responsáveis pela formulação de políticas que este relatório identifica: eles terão de ponderar frequentemente os possíveis benefícios e riscos dos avanços iminentes em inteligência artificial sem dispor de muitas evidências científicas. Ainda assim, gerar evidências sobre as implicações de segurança e proteção das tendências indicadas por o3 será uma prioridade urgente para a pesquisa em IA nas próximas semanas e meses.

Principais conclusões do relatório

● **As capacidades da IA de propósito geral, o tipo de IA em que este relatório se concentra, aumentaram rapidamente nos últimos anos e melhoraram ainda mais nos últimos**

meses. Há alguns anos, os melhores modelos de linguagem de grande porte (MLA) raramente conseguiam produzir um parágrafo de texto coerente. Atualmente, a IA de propósito geral pode escrever programas de computador, gerar imagens fotorrealistas personalizadas e participar de conversas abertas prolongadas. Desde a publicação do Relatório Provisório (maio de 2024), os novos modelos demonstraram um desempenho notavelmente superior em testes de raciocínio científico e programação.

● **Atualmente, muitas empresas estão investindo no desenvolvimento de agentes de IA de propósito geral como uma possível via para continuar avançando. Os agentes de IA são**

sistemas de IA de propósito geral que podem agir, planejar e delegar de forma autônoma para alcançar objetivos com pouca ou nenhuma supervisão humana. Os agentes de IA sofisticados poderiam, por exemplo, usar computadores para concluir projetos mais longos que os sistemas atuais, o que geraria benefícios e riscos adicionais.

● **Os avanços de capacidade que ocorrerão nos próximos meses e anos poderiam variar de lentos a extremamente rápidos. O progresso dependerá da**

capacidade das empresas de disponibilizar rapidamente ainda mais dados e poder computacional para treinar novos modelos, e da possibilidade de esse “escalonamento” dos modelos superar suas limitações atuais. Pesquisas recentes sugerem que modelos que escalam rapidamente podem continuar fisicamente viáveis por pelo menos vários anos. Mas, para alcançar avanços importantes de capacidade, talvez também sejam necessários outros fatores: por exemplo, novos avanços de pesquisa, que são difíceis de prever, ou o sucesso de uma nova abordagem de escalonamento que as empresas tenham adotado recentemente.

● **Vários dos danos que a IA de propósito geral pode causar já estão bem estabelecidos.**

Eles incluem fraudes, imagens íntimas não consensuais (IINC) e material de abuso sexual infantil (MASI), resultados tendenciosos de modelos contra determinados grupos de pessoas ou determinadas opiniões, problemas de confiabilidade e violações de privacidade. Pesquisadores desenvolveram técnicas de mitigação para esses problemas, mas, até agora, nenhuma combinação de técnicas consegue resolvê-los completamente. Desde a publicação do Relatório Provisório, formas mais sutis de viés foram reveladas por novas evidências de discriminação relacionada aos sistemas de IA de propósito geral.

● **À medida que aumentam as capacidades da IA de propósito geral, surgem**

gradualmente evidências de riscos adicionais. Eles incluem riscos como impactos em grande escala no mercado de trabalho, ataques biológicos ou invasões cibernéticas realizados com IA e a perda de controle da sociedade sobre a IA de propósito geral. A interpretação dessas evidências difere entre os especialistas:

Consulte a atualização do presidente sobre os últimos avanços em IA após a redação deste relatório.

enquanto alguns pensam que esses riscos estão a décadas de distância, outros acreditam que a IA de propósito geral poderia produzir danos em escala social nos próximos anos. Os avanços recentes nas capacidades da IA de propósito geral (particularmente em testes de raciocínio científico e programação) geraram novas evidências de riscos potenciais, como invasões cibernéticas e ataques biológicos realizados com IA, levando uma importante empresa de IA a elevar sua avaliação do risco biológico de seu melhor modelo de “baixo” para “médio”.

- **As técnicas de gestão de riscos ainda estão em seus estágios iniciais, mas o progresso é** possível. Existem vários métodos técnicos que os desenvolvedores podem empregar e os reguladores podem exigir para avaliar e reduzir os riscos da IA de propósito geral, mas todos têm limitações. Por exemplo, as técnicas atuais de interpretação para explicar por que um modelo de IA de propósito geral produziu determinado resultado continuam extremamente limitadas. No entanto, os pesquisadores estão fazendo alguns avanços para enfrentar essas limitações. Por outro lado, pesquisadores e responsáveis pela formulação de políticas tentam cada vez mais padronizar as abordagens de gestão de riscos e coordená-las em nível internacional.

- **O ritmo e a imprevisibilidade dos avanços na IA de propósito geral colocam um “dilema de evidências” para os responsáveis pela formulação de políticas. Diante dos** avanços, que às vezes são rápidos e inesperados, os responsáveis pela formulação de políticas frequentemente terão de avaliar os possíveis benefícios e riscos dos avanços iminentes em inteligência artificial sem dispor de evidências científicas suficientes. Ao fazer isso, enfrentam um dilema. Por um lado, medidas preventivas de mitigação de riscos baseadas em evidências limitadas poderiam se revelar ineficazes ou desnecessárias. Por outro, esperar por evidências mais contundentes de um risco iminente poderia deixar a sociedade despreparada ou até mesmo tornar a mitigação impossível, por exemplo, se ocorrerem grandes saltos nas capacidades da IA e nos riscos associados. Empresas e governos estão desenvolvendo sistemas de alerta precoce e estruturas de gestão de riscos que poderiam reduzir esse dilema. Algumas delas ativam medidas específicas de mitigação quando surgem novas evidências de riscos, enquanto outras exigem que os desenvolvedores forneçam evidências de segurança antes de lançar um novo modelo.

- **Há um amplo consenso entre os pesquisadores sobre a utilidade de avançar nas seguintes questões**: Com que rapidez as capacidades da IA de propósito geral avançarão nos próximos anos e como os pesquisadores podem medir esse progresso de forma confiável? Quais são os limites de risco razoáveis para acionar medidas de mitigação? Qual é a melhor maneira de os responsáveis pela formulação de políticas obterem acesso a informações sobre a IA de uso geral que sejam relevantes para a segurança pública? Como pesquisadores, empresas de tecnologia e governos podem avaliar de maneira confiável os riscos do desenvolvimento e da implantação da IA de propósito geral? Como os modelos de IA de propósito geral funcionam internamente? Como é possível projetar a IA de propósito geral para que se comporte de maneira segura?

Consulte a atualização do presidente sobre os últimos avanços em IA após a redação deste relatório.

- **A IA não é algo independente de nós: As decisões que as pessoas tomam** determinam seu futuro. O futuro da tecnologia de IA de propósito geral é incerto, com uma ampla gama de trajetórias possíveis mesmo no futuro próximo, incluindo resultados muito positivos e muito negativos. Essa incerteza pode gerar fatalismo e fazer com que a IA pareça estar fora de nosso controle. Mas serão as decisões das sociedades e dos governos sobre como lidar com essa incerteza que determinarão nosso caminho. Este relatório pretende facilitar um debate construtivo sobre essas decisões, baseado em evidências.

Consulte a atualização do presidente sobre os últimos avanços em IA após a redação deste relatório.

Resumo executivo

O propósito deste relatório

Este relatório sintetiza o estado da compreensão científica da IA de propósito geral (IA que pode realizar uma ampla variedade de tarefas), com foco na compreensão e na gestão de seus riscos.

Este relatório resume as evidências científicas sobre a segurança da IA de propósito geral. O propósito deste relatório é ajudar a criar uma compreensão internacional compartilhada dos riscos da IA avançada e de como podem ser mitigados. Para isso, concentra-se na IA, pois esse tipo de IA progrediu de forma particularmente rápida nos últimos anos e foi amplamente implementado em empresas de tecnologia para uma variedade de fins comerciais e de negócios. O relatório sintetiza o estado da compreensão científica da IA de propósito geral, com foco na compreensão e na gestão de seus riscos.

Em um contexto de rápidos avanços, a pesquisa sobre IA de propósito geral está atualmente em uma época de descobertas científicas e, em muitos casos, ainda não é uma ciência consolidada. O relatório apresenta um resumo da compreensão científica atual da IA de propósito geral e de seus riscos. Isso inclui a identificação de áreas de consenso científico e de áreas nas quais existem diferentes pontos de vista ou lacunas na compreensão científica atual.

O mundo só poderá desfrutar plenamente e com segurança dos benefícios potenciais da IA de propósito geral se seus riscos forem gerenciados adequadamente. Este relatório se concentra em identificar esses riscos e avaliar métodos técnicos para analisá-los e mitigá-los, incluindo as formas pelas quais a própria IA de propósito geral pode ser utilizada para mitigar os riscos. Não tem como finalidade avaliar exaustivamente todos os possíveis impactos sociais da IA de propósito geral. Cabe ressaltar que os benefícios atuais e futuros da IA de propósito geral, embora sejam muito amplos, ficam fora do escopo deste relatório. A formulação de políticas holísticas exige considerar tanto os benefícios potenciais da IA de propósito geral quanto os riscos abordados neste relatório. Também é necessário levar em conta que outros tipos de IA têm perfis de risco/benefício diferentes em comparação com a IA de propósito geral atual.

As três principais seções do relatório resumem as evidências científicas sobre três perguntas fundamentais: O que a IA de propósito geral pode fazer? Quais são os riscos associados à IA de propósito geral? E quais técnicas de mitigação existem contra esses riscos?

Seção 1 - Capacidades da IA de propósito geral: O que a IA de propósito geral pode fazer agora e no futuro?

As capacidades da IA de propósito geral melhoraram rapidamente nos últimos anos e os avanços futuros poderiam variar de lentos a extremamente rápidos.

Os riscos da IA de capacidade geral dependem diretamente de suas capacidades e, segundo muitas métricas, estas progrediram rapidamente. Há cinco anos, os principais modelos de linguagem de IA de propósito geral raramente conseguiam criar um parágrafo de texto coerente. Atualmente, alguns modelos de IA de propósito geral podem conversar sobre uma ampla gama de temas, escrever programas de computador ou gerar vídeos curtos realistas a partir de uma descrição. No entanto, é tecnicamente difícil estimar e descrever de forma confiável as capacidades da IA de propósito geral.

Os desenvolvedores aumentaram continuamente os recursos utilizados para treinar novos modelos (isso costuma ser chamado de "escalonamento") e aperfeiçoaram as abordagens existentes para utilizar esses recursos de maneira mais eficiente. Por exemplo, segundo estimativas recentes, os modelos de IA mais avançados tiveram aumentos anuais de aproximadamente 4 vezes nos recursos computacionais utilizados para o treinamento e de 2,5 vezes no tamanho do conjunto de dados de treinamento.

O ritmo do progresso futuro nas capacidades da IA de propósito geral tem implicações substanciais para a gestão dos riscos emergentes, mas os especialistas não chegam a um consenso sobre o que esperar mesmo nos próximos meses e anos. Segundo as diversas opiniões dos especialistas, as capacidades da IA de propósito geral podem avançar de forma lenta, rápida ou extremamente rápida.

Os especialistas não chegam a um consenso sobre o ritmo do progresso futuro devido às diferentes opiniões sobre o potencial de um maior "escalonamento", e as empresas estão explorando um novo tipo adicional de escalonamento que poderia acelerar ainda mais as capacidades. Embora o escalonamento costume superar as limitações dos sistemas anteriores, os especialistas discordam sobre seu potencial para resolver as limitações restantes dos sistemas atuais, como a falta de confiabilidade para agir no mundo físico e executar tarefas prolongadas em computadores. Nos últimos meses, um novo tipo de escalonamento demonstrou potencial para melhorar ainda mais as capacidades: além de simplesmente ampliar os recursos utilizados para treinar modelos, as empresas de IA também estão cada vez mais interessadas no "escalonamento de inferência". Ou seja, permitir que um modelo já treinado use mais computação para resolver um problema determinado, por exemplo, para melhorar sua própria solução ou escrever as chamadas "cadeias de pensamento", que dividem o problema em etapas mais simples.

Consulte a atualização do presidente sobre os últimos avanços em IA após a redação deste relatório.

Várias empresas líderes que desenvolvem IA de propósito geral estão apostando no "escalonamento" para continuar impulsionando melhorias de desempenho. Se as tendências recentes se mantiverem, até o fim de 2026 alguns modelos de IA de propósito geral serão treinados utilizando aproximadamente 100 vezes mais recursos computacionais que os modelos mais intensivos de 2023. Isso levaria ao uso de 10 000 vezes mais recursos computacionais de treinamento até 2030, combinado com algoritmos que alcançam maiores capacidades para uma determinada quantidade de computação disponível. Além desse potencial escalonamento dos recursos de treinamento, tendências recentes, como o escalonamento de inferência e o uso de modelos para gerar dados de treinamento, poderiam significar que, no conjunto, se utilizará ainda mais computação. No entanto, existem possíveis obstáculos que impedem continuar aumentando rapidamente os dados e os recursos computacionais, como a disponibilidade de dados, os chips de IA, o capital e a capacidade energética local. As empresas que desenvolvem IA de propósito geral estão trabalhando para superar esses possíveis obstáculos.

Desde a publicação do Relatório Provisório (maio de 2024), a IA de propósito geral alcançou um desempenho de nível especializado em alguns testes e competições de raciocínio científico e programação, e as empresas vêm fazendo grandes esforços para desenvolver agentes autônomos de IA. Os avanços em ciência e programação foram impulsionados por técnicas de escalonamento de inferência, como a escrita de longas "cadeias de pensamento". Novos estudos sugerem que ampliar ainda mais essas abordagens, por exemplo, permitindo que os modelos analisem problemas escrevendo cadeias de pensamento ainda mais longas que as dos modelos atuais, poderia levar a maiores avanços em domínios nos quais o raciocínio é mais importante, como a ciência, a engenharia de software e o planejamento. Além dessa tendência, as empresas estão fazendo grandes esforços para desenvolver agentes de IA de propósito geral mais avançados, capazes de planejar e agir de forma autônoma para trabalhar em direção a um objetivo determinado. Por fim, o preço de mercado do uso de um sistema de IA de propósito geral com determinado nível de capacidade caiu drasticamente, tornando essa tecnologia mais acessível e mais amplamente utilizada.

Este relatório se concentra principalmente nos aspectos técnicos do progresso da IA, mas a velocidade com que a IA de propósito geral avançará não é uma questão puramente técnica. O ritmo dos avanços futuros também dependerá de fatores não técnicos, incluindo potencialmente as abordagens que os governos adotarem para regular a IA. Este relatório não analisa como as diferentes abordagens de regulação poderiam afetar a velocidade de desenvolvimento e adoção da IA de propósito geral.

Seção 2 - Riscos: Quais são os riscos associados à IA de propósito geral?

Vários danos que a IA de propósito geral pode causar já estão estabelecidos. À medida que aumentam as capacidades da IA de propósito geral, surgem gradualmente evidências de riscos adicionais.

Este relatório classifica os riscos da IA de propósito geral em três categorias: riscos de uso malicioso, riscos decorrentes de mau funcionamento e riscos sistêmicos. Cada uma dessas categorias contém riscos que já se materializaram, assim como riscos que poderiam se materializar nos próximos anos.

***Riscos de uso mal-intencionado*: agentes maliciosos podem usar a IA de propósito geral para causar danos a indivíduos, organizações ou à sociedade. As formas de uso malicioso** incluem:

- **Danos a indivíduos por meio de conteúdo falso**: Atualmente, agentes maliciosos podem usar a IA de propósito geral para gerar conteúdo falso que prejudique pessoas de maneira direcionada. Esses usos maliciosos incluem pornografia “deepfake” não consensual e MASI gerado por IA, fraude financeira por imitação de voz, chantagem para extorsão, sabotagem da reputação pessoal e profissional e abuso psicológico. No entanto, embora sejam comuns os relatos de incidentes de danos causados por conteúdo falso gerado por IA, ainda faltam estatísticas confiáveis sobre a frequência desses incidentes.
- Manipulação da opinião pública: A IA de propósito geral facilita a geração de conteúdo persuasivo em grande escala. Isso pode ajudar agentes que buscam manipular a opinião pública, por exemplo, para afetar resultados políticos. No entanto, as evidências sobre a frequência e a eficácia desses esforços continuam limitadas. Contramedidas técnicas, como marcas d’água no conteúdo, apesar de úteis, geralmente podem ser contornadas por agentes moderadamente sofisticados.
- Crimes cibernéticos: A IA de propósito geral pode facilitar ou acelerar a execução de ataques cibernéticos por agentes maliciosos com diferentes níveis de habilidade. Os sistemas atuais demonstraram capacidades em tarefas de cibersegurança de complexidade baixa e média, e agentes patrocinados por um Estado estão explorando ativamente a IA para inspecionar sistemas-alvo. Uma nova pesquisa confirmou que as capacidades da IA de propósito geral relacionadas aos crimes cibernéticos estão avançando significativamente, mas ainda não está claro se isso afetará o equilíbrio entre atacantes e defensores.

● **Ataques biológicos e químicos**: Sistemas recentes de IA de propósito geral demonstraram certa capacidade de fornecer instruções e orientações para solucionar problemas com a finalidade de reproduzir armas biológicas e químicas conhecidas e facilitar o projeto de novos compostos tóxicos. Em novos experimentos que testaram a capacidade de gerar planos para produzir armas biológicas, um sistema de IA de propósito geral obteve, em algumas ocasiões, resultados melhores que os de especialistas humanos com acesso à Internet. Em resposta a isso, uma empresa de IA elevou a avaliação do risco biológico de seu melhor modelo de “baixo” para “médio”. No entanto, as tentativas reais de desenvolver essas armas ainda exigem recursos e conhecimentos adicionais. É complexo realizar uma avaliação abrangente do risco biológico e químico porque grande parte da pesquisa pertinente é sigilosa.

Desde a publicação do Relatório Provisório (maio de 2024), a capacidade da IA de propósito geral aumentou em domínios relevantes para o uso malicioso. Por exemplo, pesquisadores construíram recentemente sistemas de IA de propósito geral capazes de encontrar e explorar por conta própria algumas vulnerabilidades de cibersegurança e, com assistência humana, descobrir uma vulnerabilidade antes desconhecida em um software amplamente utilizado. As capacidades da IA de propósito geral também melhoraram em relação ao raciocínio e à integração de diferentes tipos de dados, o que auxilia a pesquisa sobre patógenos ou outros campos de uso duplo.

***Riscos de falhas*: A IA de propósito geral também pode causar danos não intencionais. Mesmo** quando os usuários não têm a intenção de causar danos, podem surgir riscos graves devido ao mau funcionamento da IA de propósito geral. Entre esses maus funcionamentos estão os seguintes:

● **Problemas de confiabilidade**: A IA de propósito geral atual pode ser pouco confiável, o que pode provocar danos. Por exemplo, se os usuários consultarem um sistema de IA de propósito geral para obter orientação médica ou jurídica, o sistema poderia gerar uma resposta com informações falsas. Esses usuários geralmente não têm consciência das limitações de um produto de IA, por exemplo, devido a uma alfabetização limitada em IA, publicidade enganosa ou erros de comunicação. Existem numerosos casos de danos decorrentes de problemas de confiabilidade, mas as evidências sobre a magnitude desses tipos de problema continuam limitadas.
● Vieses: A IA de propósito geral pode amplificar os vieses sociais e políticos, causando danos concretos. Esses sistemas costumam apresentar vieses em relação à raça, ao gênero, à cultura, à idade, às deficiências, à opinião política ou a outros aspectos da identidade humana. Isso pode gerar resultados discriminatórios, incluindo a alocação desigual de recursos, o reforço de estereótipos e a rejeição sistemática de grupos ou pontos de vista sub-representados. As abordagens técnicas para mitigar o viés e a discriminação nos sistemas de IA de propósito geral estão avançando, mas enfrentam uma escolha difícil entre a

mitigação do viés e objetivos concorrentes, como a precisão e a privacidade, além de outros desafios.

● **Perda de controle**: Os cenários de "perda de controle" são cenários futuros hipotéticos nos quais um ou mais sistemas de IA de propósito geral começam a operar fora do controle de qualquer pessoa, sem um caminho claro para recuperá-lo. Há um amplo consenso de que a IA de propósito geral atualmente não apresenta esse risco por não dispor das capacidades necessárias. No entanto, a opinião dos especialistas sobre a probabilidade de perda de controle nos próximos anos varia enormemente: alguns a consideram inverossímil, outros consideram provável que ocorra e outros a veem como um risco de probabilidade moderada que merece atenção devido à sua alta gravidade potencial. As pesquisas empíricas e matemáticas em andamento estão fazendo esse debate avançar gradualmente.

Desde a publicação do Relatório Provisório (maio de 2024), novas pesquisas trouxeram conhecimentos inéditos sobre os riscos de viés e perda de controle. As
evidências de vieses nos sistemas de IA de propósito geral aumentaram e trabalhos de pesquisa recentes detectaram formas adicionais de vieses em IA. Os pesquisadores observaram modestos avanços adicionais rumo às capacidades de IA que provavelmente seriam necessárias para que ocorram os cenários de perda de controle comumente discutidos. Elas incluem capacidades para utilizar computadores de forma autônoma, programar, obter acesso não autorizado a sistemas digitais e identificar formas de contornar a supervisão humana.

***Riscos sistêmicos*: além dos riscos diretamente associados às capacidades de modelos individuais, a implantação generalizada da IA de propósito geral está relacionada**
a vários riscos sistêmicos mais amplos. Os exemplos de riscos sistêmicos vão desde possíveis impactos no mercado de trabalho até riscos de privacidade e efeitos ambientais:

● **Riscos no mercado de trabalho**: A IA de propósito geral (em particular, se continuar avançando rapidamente) tem o potencial de automatizar uma ampla gama de tarefas, o que poderia ter um efeito significativo no mercado de trabalho. Isso significa que muitas pessoas poderiam perder seus empregos atuais. No entanto, muitos economistas esperam que as possíveis perdas de empregos possam ser compensadas, em parte ou até totalmente, pela criação de novos empregos e por uma maior demanda em setores não automatizados.

● Lacuna global em P&D de IA: Atualmente, a pesquisa e o desenvolvimento em IA de propósito geral se concentram em poucos países ocidentais e na China. Essa “lacuna da IA” tem o potencial de aumentar a dependência de grande parte do mundo em relação a esse pequeno conjunto de países. Alguns especialistas também esperam que isso contribua para a desigualdade global. A lacuna tem muitas causas, incluindo algumas que não são exclusivas da IA. No entanto, isso se deve, em grande medida, aos diferentes níveis de acesso aos

dispendiosos recursos computacionais necessários para desenvolver sistemas de IA de propósito geral: A maioria dos países de renda baixa e média (PIBM) tem consideravelmente menos acesso a recursos computacionais que os países de renda alta (PIA).

● Concentração de mercado e pontos únicos de falha: Atualmente, um pequeno número de empresas domina o mercado de IA de propósito geral. Essa concentração de mercado poderia tornar as sociedades mais vulneráveis a vários riscos sistêmicos. Por exemplo, se todas as organizações de setores críticos, como finanças ou saúde, dependerem de uma pequena quantidade de sistemas de IA de propósito geral, um erro ou uma vulnerabilidade em um desses sistemas poderia causar falhas e interrupções simultâneas em grande escala.

● Riscos ambientais: O uso crescente de recursos computacionais no desenvolvimento e na implantação da IA de propósito geral aumentou rapidamente as quantidades de energia, água e matérias-primas consumidas na construção e no funcionamento da infraestrutura computacional necessária. Apesar do progresso nas técnicas que permitem usar recursos computacionais de maneira mais eficiente, essa tendência não mostra sinais definitivos de desaceleração. A IA de propósito geral também tem uma série de aplicações que podem beneficiar ou prejudicar as iniciativas de sustentabilidade.

● Riscos de privacidade: A IA de propósito geral pode causar ou contribuir para violações da privacidade do usuário. Por exemplo, informações confidenciais presentes nos dados de treinamento podem vazar involuntariamente quando um usuário interage com o sistema. Além disso, quando os usuários compartilham informações confidenciais com o sistema, essas informações também podem vazar. A IA de propósito geral também pode facilitar violações deliberadas de privacidade, por exemplo, se agentes maliciosos a utilizarem para inferir informações sensíveis sobre indivíduos específicos a partir de grandes quantidades de dados. No entanto, até agora os pesquisadores não encontraram evidências de violações generalizadas de privacidade associadas à IA de propósito geral.

● Violações de direitos autorais: A IA de propósito geral aprende com obras de expressão criativa e também as cria, desafiando os sistemas tradicionais de consentimento, compensação e controle de dados. A coleta de dados e a geração de conteúdo podem envolver uma variedade de leis sobre direitos relativos a dados, que variam conforme a jurisdição e podem ser objeto de litígios em andamento. Diante da incerteza jurídica em torno das práticas de coleta de dados, as empresas de IA estão compartilhando menos informações sobre os dados que utilizam. Essa opacidade dificulta a pesquisa sobre a segurança da IA por terceiros.

Desde a publicação do Relatório Provisório (maio de 2024), surgiram evidências adicionais sobre os impactos da IA de propósito geral no mercado de trabalho, enquanto os novos avanços alimentaram as preocupações com a privacidade e os direitos
autorais. As novas análises dos dados do mercado de trabalho sugerem que as pessoas estão

adotando a IA de propósito geral muito rapidamente em comparação com as tecnologias anteriores. O ritmo de adoção pelas empresas varia amplamente conforme o setor. Além disso, os avanços recentes nas capacidades ajudaram a fazer com que a IA de propósito geral seja implementada cada vez mais em contextos sensíveis, como a saúde ou o monitoramento do local de trabalho, criando novos riscos para a privacidade. Por fim, com a intensificação das disputas sobre direitos autorais e a persistente falta de confiabilidade das medidas técnicas para mitigar violações desses direitos, os titulares de direitos sobre dados restringiram rapidamente o acesso a seus dados.

***Modelos de pesos abertos*: um fator importante ao avaliar muitos dos riscos que um modelo de IA de propósito geral poderia apresentar é como ele é disponibilizado ao público. Os** chamados "modelos de pesos abertos" são modelos de IA cujos componentes centrais, chamados "pesos", são compartilhados publicamente para download. O acesso aos pesos abertos facilita a pesquisa e a inovação, inclusive em segurança da IA, além de aumentar a transparência e ajudar a comunidade de pesquisa a detectar falhas nos modelos. No entanto, modelos de pesos abertos também podem apresentar riscos, por exemplo, ao facilitar usos mal-intencionados ou equivocados que sejam difíceis ou impossíveis de controlar pelo desenvolvedor do modelo. Uma vez que os pesos do modelo estejam disponíveis para download público, não há como reverter de forma generalizada todas as cópias existentes, nem garantir que todas recebam atualizações de segurança. Desde o Relatório Provisório (maio de 2024), surgiu um consenso de alto nível: propõe-se avaliar os riscos que uma maior abertura da IA apresenta em termos de risco “marginal”. Ou seja, a medida em que o lançamento de um modelo de pesos abertos aumentaria ou diminuiria determinado risco em relação aos riscos apresentados por alternativas existentes, como modelos fechados ou outras tecnologias.

Seção 3 - Gestão de riscos: Quais técnicas existem para gerenciar os riscos da IA de propósito geral?

Existem várias abordagens técnicas que podem ajudar a gerenciar os riscos, mas, em muitos casos, as melhores abordagens disponíveis ainda têm limitações muito significativas e não oferecem estimativas quantitativas dos riscos nem garantias que existem em outros domínios críticos para a segurança.

A gestão de riscos (identificar e avaliar os riscos e, depois, mitigá-los e controlá-los) é difícil no contexto da IA de propósito geral. Embora a gestão de riscos também tenha sido um grande desafio em muitos outros domínios, algumas características da IA de propósito geral parecem criar dificuldades específicas.

Várias características técnicas da IA de propósito geral tornam a gestão de riscos particularmente difícil nesse âmbito. Essas características incluem, entre outras, as seguintes:

- **A gama de possíveis usos e contextos de uso dos sistemas de IA de propósito geral** é excepcionalmente ampla. Por exemplo, o mesmo sistema pode ser utilizado para fornecer orientação médica, analisar códigos de computador para detectar vulnerabilidades e gerar fotografias. Isso aumenta a dificuldade de antecipar de forma abrangente os casos de uso relevantes, identificar riscos ou testar como os sistemas se comportarão em circunstâncias relevantes do mundo real.
- **Os desenvolvedores ainda entendem pouco sobre como funcionam seus modelos de IA de** propósito geral. Essa falta de compreensão torna mais difícil prever problemas de comportamento, assim como explicar e resolver problemas conhecidos depois de observados. A compreensão continua difícil principalmente porque os modelos de IA de propósito geral não são programados no sentido tradicional. Em vez disso, são treinados: os desenvolvedores de IA configuram um processo de treinamento que envolve um grande volume de dados, e o resultado desse processo é o modelo de IA de propósito geral. O funcionamento interno desses modelos é, em grande medida, incompreensível, inclusive para seus desenvolvedores. As técnicas de explicação dos modelos e de "interpretabilidade" podem melhorar a compreensão de pesquisadores e desenvolvedores sobre como os modelos de IA de propósito geral funcionam, mas, apesar dos avanços recentes, essa pesquisa ainda está em seus estágios iniciais.

- **Os agentes de IA (sistemas de IA de propósito geral que podem agir, planejar e delegar de forma autônoma para alcançar objetivos), cada vez mais capazes, provavelmente apresentarão novos e consideráveis desafios para a gestão de riscos.**
  Os agentes de IA normalmente trabalham para alcançar objetivos de forma autônoma utilizando software de uso geral, como navegadores web e ferramentas de programação. Atualmente, a maioria ainda não é suficientemente confiável para uso generalizado, mas as empresas estão se esforçando para construir agentes de IA mais capazes e confiáveis e fizeram avanços nos últimos meses. É provável que os agentes de IA se tornem cada vez mais úteis, mas eles também podem agravar vários dos riscos analisados neste relatório e introduzir dificuldades adicionais para a gestão de riscos. Alguns exemplos desses novos desafios potenciais incluem a possibilidade de os usuários nem sempre saberem o que seus próprios agentes de IA estão fazendo, a possibilidade de os agentes de IA operarem fora do controle de qualquer pessoa, a possibilidade de atacantes assumirem o controle de outros agentes e a possibilidade de as interações entre diferentes sistemas de IA criarem novos riscos complexos. As abordagens para gerenciar os riscos associados aos agentes estão apenas começando a ser desenvolvidas.

Além dos fatores técnicos, vários fatores econômicos, políticos e sociais tornam a gestão de riscos no campo da IA de propósito geral particularmente difícil.

. Consulte a atualização do presidente sobre os últimos avanços em IA após a redação deste relatório.

- **O ritmo do avanço da IA de propósito geral cria um "dilema de evidências" para os** tomadores de decisão. O rápido avanço das capacidades faz com que alguns riscos surjam em grandes saltos; por exemplo, o risco de fraude acadêmica com o uso da IA de propósito geral passou de insignificante a generalizado no prazo de um ano. Quanto mais rapidamente um risco surge, mais difícil é gerenciá-lo de forma reativa e mais valiosa se torna a preparação. No entanto, enquanto as evidências de um risco permanecerem incompletas, os tomadores de decisão também não poderão saber com certeza se o risco surgirá ou mesmo se já surgiu. Isso cria uma escolha difícil: implementar medidas preventivas ou de mitigação precoce pode se revelar desnecessário, mas esperar por evidências conclusivas poderia deixar a sociedade vulnerável a riscos que surgem rapidamente. Empresas e governos estão desenvolvendo sistemas de alerta precoce e estruturas de gestão de riscos que podem reduzir esse dilema. Alguns deles ativam medidas específicas de mitigação quando surgem novas evidências de riscos, enquanto outros exigem que os desenvolvedores forneçam evidências de segurança antes de lançar um novo modelo.

- **Existe uma lacuna de informação entre o que as empresas de IA sabem sobre seus sistemas de IA e o que os governos e os pesquisadores de fora da indústria sabem. As** empresas costumam compartilhar informações limitadas sobre seus sistemas de IA de propósito geral, especialmente no período anterior ao lançamento para o público em geral. As empresas citam diversas preocupações comerciais e de segurança como razões para limitar o compartilhamento de informações. No entanto, essa lacuna de informação também dificulta a participação efetiva de outros agentes na gestão de riscos, especialmente no caso de riscos emergentes.
- **Tanto as empresas de IA quanto os governos costumam enfrentar uma forte pressão competitiva, o que pode fazer com que não priorizem a gestão de riscos. Em algumas** circunstâncias, a pressão competitiva pode incentivar as empresas a investir menos tempo ou outros recursos na gestão de riscos do que investiriam em outras circunstâncias. Da mesma forma, os governos podem investir menos em políticas de apoio à gestão de riscos quando percebem uma escolha difícil entre a competição internacional e a redução de riscos.

Ainda assim, existem diversas técnicas e estruturas para gerenciar os riscos decorrentes da IA de propósito geral que as empresas podem empregar e os reguladores podem exigir. Elas incluem métodos para identificar e avaliar riscos, assim como métodos para mitigá-los e controlá-los.

- **Avaliar os sistemas de IA de propósito geral quanto a riscos é parte integrante da gestão de riscos, mas as avaliações de risco existentes são extremamente** limitadas. As avaliações existentes dos riscos da IA de propósito geral se baseiam principalmente em "verificações pontuais", ou seja, em testar o comportamento da IA de propósito geral em um conjunto de situações específicas. Isso pode ajudar a detectar perigos potenciais antes de implantar um modelo. No entanto, os testes existentes muitas

vezes deixam passar perigos e superestimam ou subestimam as capacidades e os riscos da IA de propósito geral, pois as condições de teste diferem do mundo real.

● **Para que a identificação e a avaliação de riscos sejam eficazes, os avaliadores precisam de considerável experiência, recursos e acesso suficiente às informações relevantes. Uma** avaliação rigorosa de riscos no contexto da IA de propósito geral exige a combinação de múltiplas abordagens de avaliação. Elas vão desde análises técnicas dos modelos e sistemas até avaliações de possíveis riscos decorrentes de determinados padrões de uso. Os avaliadores precisam de experiência considerável para realizar essas avaliações corretamente. Para realizar avaliações abrangentes de riscos, também costumam precisar de mais tempo, de acesso mais direto aos modelos e a seus dados de treinamento e de mais informações sobre as metodologias técnicas utilizadas do que as empresas que desenvolvem IA de propósito geral geralmente fornecem.

● **Houve avanços no treinamento de modelos de IA de propósito geral para que funcionem de maneira mais segura, mas nenhum método atual consegue prevenir de forma** confiável nem mesmo saídas abertamente inseguras. Por exemplo, uma técnica chamada "treinamento adversarial" envolve expor deliberadamente modelos de IA a exemplos projetados para fazê-los falhar ou funcionar mal durante o treinamento, com o objetivo de gerar resistência a esses casos. No entanto, os adversários ainda conseguem encontrar novas formas ("ataques") de contornar essas medidas de segurança com um nível baixo ou moderado de esforço. Além disso, evidências recentes sugerem que os métodos atuais de treinamento, que dependem em grande medida de um feedback humano imperfeito, podem ajudar inadvertidamente os modelos a enganar os humanos em questões complexas, tornando os erros mais difíceis de detectar. Melhorar a quantidade e a qualidade desse feedback é uma forma de avançar, embora técnicas de treinamento emergentes que utilizam IA para detectar comportamentos enganosos também apresentem resultados promissores.

● **O monitoramento, que consiste em identificar riscos e avaliar o desempenho depois que um modelo já está em uso, e diversas intervenções para prevenir ações prejudiciais podem melhorar a segurança de uma IA de propósito geral depois de ela ter sido** disponibilizada aos usuários. As ferramentas atuais podem detectar conteúdo gerado por IA, acompanhar o desempenho do sistema e identificar entradas/saídas potencialmente prejudiciais, embora usuários moderadamente capacitados costumem contornar essas proteções. Várias camadas de defesa que combinam capacidades de intervenção e controle técnico com supervisão humana melhoram a segurança, mas podem causar custos e atrasos. No futuro, mecanismos habilitados por hardware poderiam ajudar clientes e reguladores a controlar os sistemas de IA de propósito geral de maneira mais eficaz durante a implantação e, potencialmente, ajudar a verificar acordos entre países, mas ainda não existem mecanismos confiáveis desse tipo.

● **Existem múltiplos métodos ao longo do ciclo de vida da IA para proteger a** privacidade. Eles incluem a remoção de informações confidenciais dos dados de treinamento, abordagens de treinamento de modelos que controlam quanta informação

é aprendida com os dados (como as abordagens de “privacidade diferencial”) e técnicas para usar IA com dados sensíveis que dificultam a recuperação dos dados (como a “computação confidencial” e outras tecnologias que melhoram a privacidade). Muitos métodos de melhoria da privacidade de outros campos de pesquisa ainda não são aplicáveis aos sistemas de IA de propósito geral devido aos requisitos computacionais desses sistemas. Nos últimos meses, os métodos de proteção da privacidade se ampliaram para abordar o uso crescente da IA em domínios sensíveis, incluindo assistentes de smartphones, agentes de IA, assistentes de voz que estão sempre ouvindo e o uso na saúde ou na prática jurídica.

Desde a publicação do Relatório Provisório (maio de 2024), os pesquisadores alcançaram avanços adicionais para explicar por que um modelo de IA de propósito geral
produz determinado resultado. Poder explicar as decisões da IA poderia ajudar a gerenciar os riscos decorrentes de mau funcionamento, que vão de vieses e inexatidões factuais à perda de controle. Além disso, recentemente foram realizadas iniciativas para padronizar as abordagens de avaliação e mitigação em todo o mundo.

Conclusão: Existe uma ampla gama de trajetórias para o futuro da IA de propósito geral e muito dependerá de como as sociedades e os governos agirem

O futuro da IA de propósito geral é incerto, com uma ampla gama de trajetórias possíveis mesmo no futuro próximo, incluindo resultados muito positivos e muito negativos. Mas nada sobre
o futuro da IA de propósito geral é inevitável. A forma como a IA de propósito geral é desenvolvida e quem a desenvolve, quais problemas ela é projetada para resolver, se as sociedades poderão aproveitar todo o potencial econômico da IA de propósito geral, quem se beneficia dela, os tipos de riscos aos quais nos expomos e quanto investimos em pesquisa para gerenciar os riscos: essas e muitas outras questões dependem das decisões que as sociedades e os governos tomarem hoje e no futuro para moldar o desenvolvimento da IA de propósito geral.

Para ajudar a facilitar um debate construtivo sobre essas decisões, este relatório apresenta uma visão geral do estado atual da pesquisa científica e do debate sobre a gestão dos riscos da IA de propósito geral. As implicações são grandes. Esperamos dar continuidade a esse esforço.