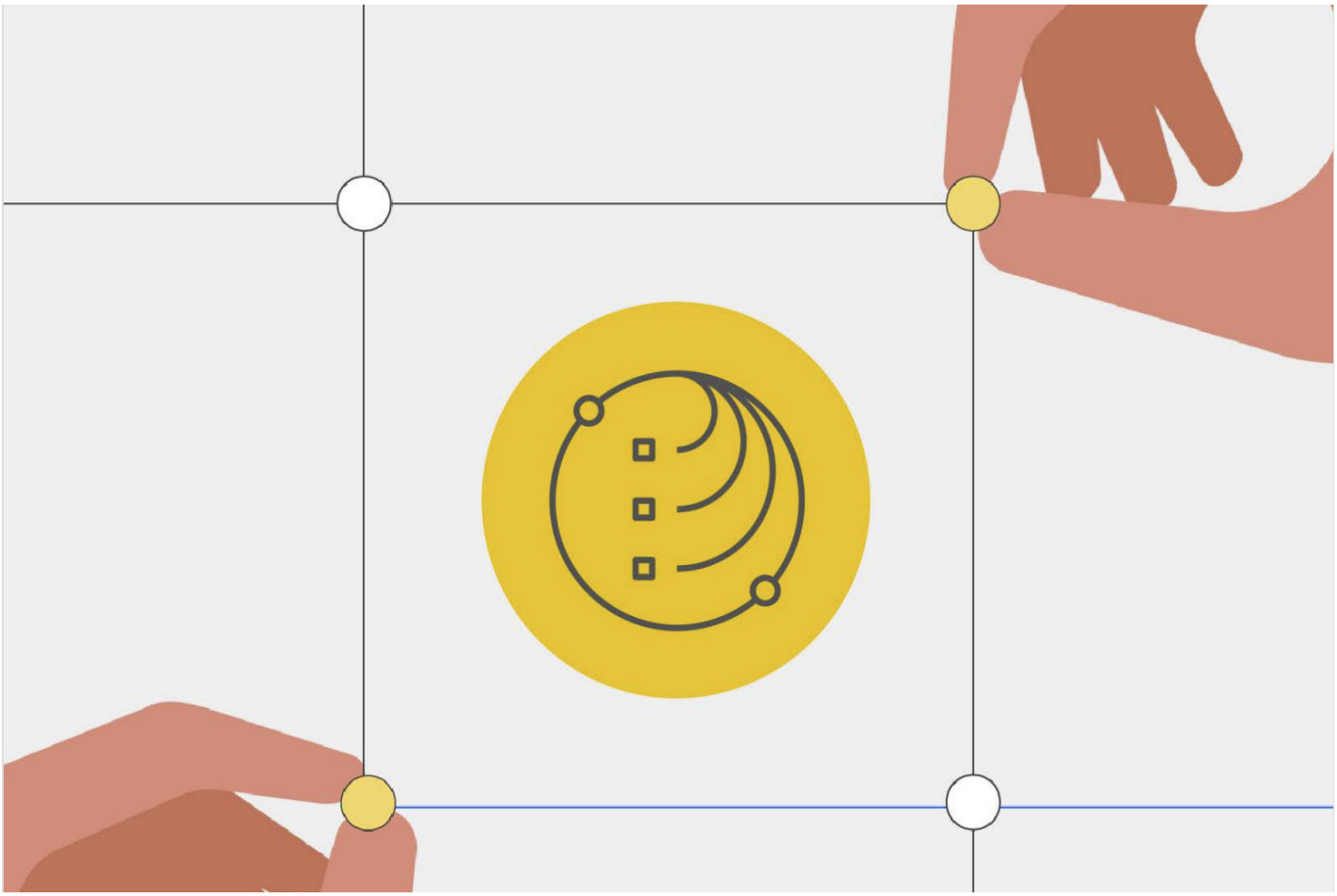
Agentes de IA:
Oportunidades, riscos e mitigações



Atribuição

Com gratidão aos:

Copresidentes do Conselho de Ética em IA e patrocinadores executivos do grupo de trabalho: Christina Montgomery, Francesca Rossi, Kush Varshney, Manish Bhide e Rob Parkin

Colaboradores: Saishruthi Swaminathan, Ambrish Rawat, Chan Nasseb, Christopher Noessel, Daniel Karl I. Weidele, David Piorkowski, Heloísa Candello, Jamie VanDodick, John Richards, Matt Bellio, Michael Hind, Michael Muller, Mihaela Bornea, Milena Pribic, Phaedra Boinodris, Rogério Abreu de Paula, Sara E Berger e Simon Rogers

Agradecimentos

Os copresidentes, patrocinadores executivos e colaboradores gostariam de agradecer aos seguintes membros pela revisão e feedback: Alina Glaubitz, Amit Dhurandhar, Christopher Hay, Jill Maguire, Katherine Fick, Kevin Black, Lauren Quigley, Manish Goyal, Maryam Ashoori, Monica Patel, Pin-Yu Chen, Ryan Hagemann e Wouter Oosterbosch

Índice

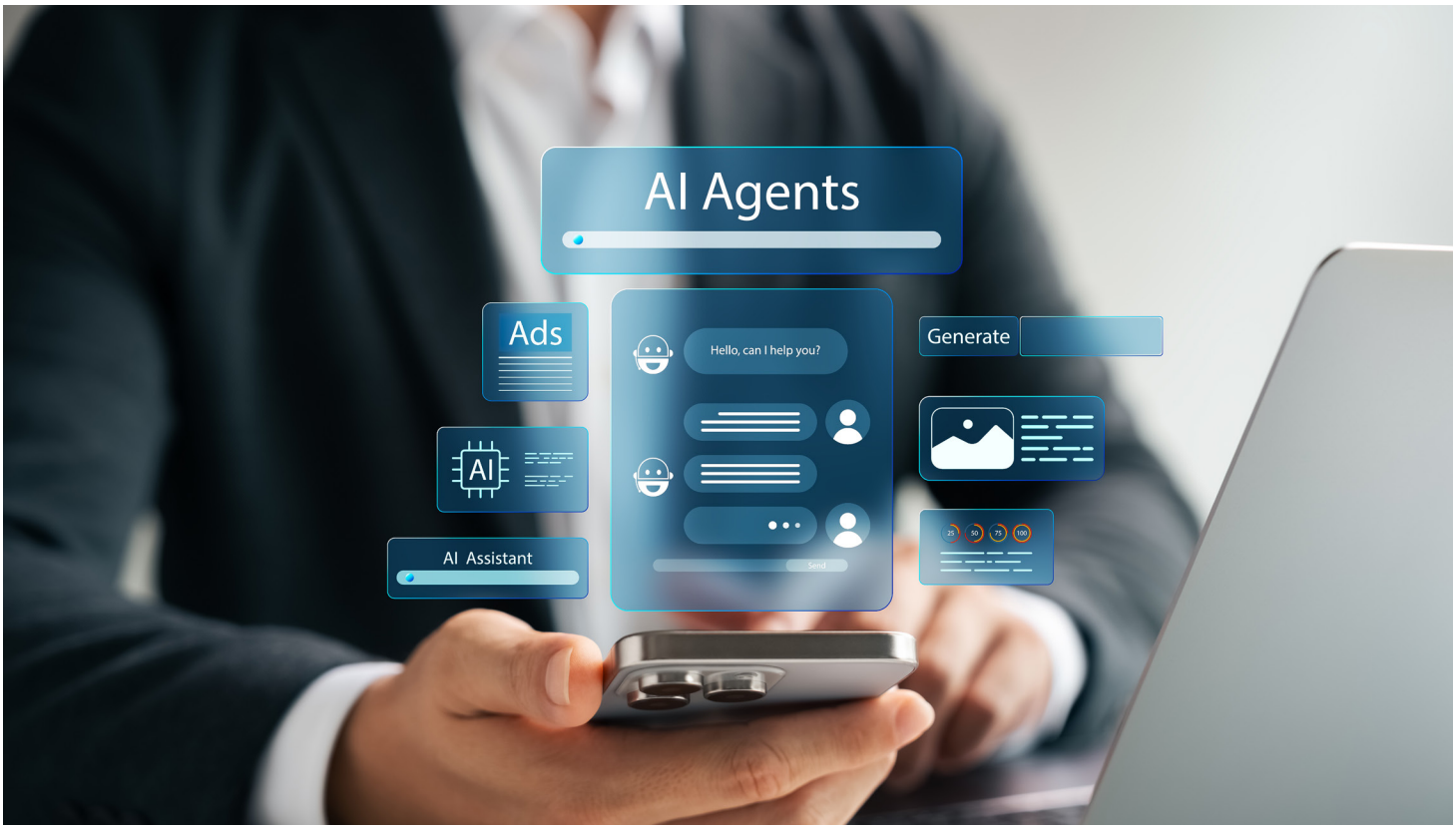
Introdução

A inteligência artificial (IA) está revolucionando a forma como as pessoas interagem com o mundo ao seu redor. À medida que a tecnologia de IA continua evoluindo, ela libera novas oportunidades e ajuda a resolver problemas complexos em diversos campos como finanças, saúde, meio ambiente, educação e esportes. Estamos agora entrando em uma era da IA em que as empresas estão adotando agentes de IA para transformar seus processos, maximizar o impacto nos negócios em várias funções e acelerar o tempo para gerar valor. Um agente de IA é uma entidade de software que emprega técnicas de IA e possui agência para agir em seu ambiente com base em metas definidas, o que significa que pode decidir quais ações executar e tem a capacidade de realizá-las. Agentes de IA existem há muito tempo, começando com agentes baseados em regras até os mais recentes agentes com grandes modelos de linguagem (LLMs), que empregam modelos de linguagem extensos e demonstram grande potencial em diversos setores. Sistemas de IA agêntica são sistemas de software que utilizam agentes de IA (junto com outros componentes como ferramentas, planejadores, memória e conjuntos de dados), perseguem objetivos e podem operar de forma autônoma. Embora agentes de IA e sistemas de IA agêntica sejam software, eles podem ser usados para controlar hardware.

***Observação**: neste artigo, os termos agente de IA e sistema de IA agêntica são usados de forma intercambiável*

A IBM é uma empresa de nuvem híbrida e IA e, por meio da força de nossas equipes de pesquisa, produto e consultoria, juntamente com parceiros externos e o Conselho de Ética em IA, ajuda a levar o poder dos agentes de IA aos nossos clientes. Alguns exemplos incluem o Agente de Engenharia de software da IBM® Research, serviços de integração de IA da IBM® Consulting, Granite BeeAI Agent Framework, watsonx Orchestrate® e watsonx.ai® da IBM Technology.

Com os rápidos avanços na tecnologia de IA, os agentes de IA estão se tornando mais poderosos e sofisticados, levantando questões sobre suas implicações éticas e impacto social. Este documento descreve o ponto de vista atual da IBM sobre a ética e a governança de agentes de IA. Esta é a versão um, e versões futuras expandirão vários aspectos da abordagem da IBM quanto à ética e à governança de seus agentes de IA.



Benefícios

Agentes de IA podem aprimorar significativamente a eficácia com que os humanos realizam tarefas e alcançam resultados de negócios. Esses benefícios incluem:

Aumento da inteligência humana

Agentes de IA podem ser integrados a fluxos de trabalho para ajudar a reduzir o tempo gasto na realização de tarefas e aprimorar o desempenho humano. Por exemplo, o sistema IBM AI Agent SWE-1.0 consiste em um agente de localização e um agente de edição que se integram aos fluxos de trabalho do GitHub para ajudar desenvolvedores a reduzir o tempo gasto encontrando bugs, desenvolvendo e testando correções para defeitos de software.

Automação

Agentes de IA podem automatizar tarefas rotineiras ou demoradas para permitir maior foco em inovação e trabalho estratégico. Por exemplo, o Assistente digital de RH AskHR da IBM usa agentes de IA para automatizar processos comuns de RH, como suporte ao funcionário e integração. Os agentes são construídos sobre o watsonx Orchestrate e impulsionados por IA generativa. Com essa automação, o AskHR da IBM agora lida com 94% das consultas de funcionários e resolve cerca de 10,1 milhões de interações por ano, permitindo que a equipe de RH da IBM se concentre em trabalhos mais importantes, como o planejamento estratégico.

Maior eficiência e produtividade

Agentes de IA podem operar continuamente, gerenciar várias tarefas simultaneamente e enfrentar desafios complexos. Isso pode ajudar a acelerar a entrega, aumentar a produtividade e melhorar a eficiência das operações de negócios. Por exemplo, a IBM está expandindo sua parceria com a Salesforce para fornecer agentes de IA pré-configurados que estarão disponíveis aos clientes 24 horas por dia, apoiando vendas e serviços. Utilizando o watsonx Orchestrate, a IBM criará agentes de IA para o Agentforce, o pacote de agentes autônomos da Salesforce, para ajudar empresas a melhorar a produtividade, mantendo a segurança.

Tomada de decisão aprimorada e qualidade das respostas

Agentes de IA podem ser conectados a vários recursos externos, ferramentas e outros agentes para ajudar a melhorar sua tomada de decisão e a qualidade de suas respostas. Eles também podem fornecer respostas abrangentes e personalizadas ao usuário, resultando em uma melhor experiência do cliente. Por exemplo, a IBM Consulting trabalhou com uma empresa global de ciências da vida para reunir uma série de agentes de IA e acelerar a geração de documentação técnica e rastreável baseada em fatos.

Riscos, desafios e impactos sociais

Como toda tecnologia em rápida evolução, agentes de IA podem criar riscos além dos benefícios. Diversas ações legais e regulatórias podem estar associadas a esses riscos, assim como consequências reputacionais e operacionais. Em geral, os riscos levantam questões sociotécnicas e devem ser tratados e mitigados por meio de métodos sociotécnicos, incluindo ferramentas de software, processos de avaliação de risco, frameworks éticos de IA, mecanismos de governança, consultas com múltiplos stakeholders, normas e regulamentações.

Abordagem

Agentes de IA podem executar três tipos de ações:

- Realizar ações que impactam o mundo (físico ou digital).
- Consultar recursos e usar ferramentas.
- Decidir qual processo escolher na seleção de recursos/ferramentas/outros agentes de IA e selecioná-los.

Agentes de IA podem realizar essas ações de forma autônoma, ou seja, sem supervisão humana contínua.

Devido à sua natureza, agentes de IA e sistemas de IA agêntica podem apresentar as seguintes características:

- Opacidade, devido à visibilidade limitada de como os agentes de IA operam, incluindo seu funcionamento interno e interações.
- Abertura, na escolha de recursos/ferramentas/outros agentes de IA para execução de ações. Isso pode aumentar a possibilidade de execução de ações inesperadas.
- Complexidade, que surge como consequência da abertura e se intensifica com a ampliação dessa abertura.
- Irreversibilidade, como consequência de ações que podem impactar o mundo.

Essas características, bem como a variedade de níveis de autonomia que agentes de IA e sistemas de IA agêntica podem ter, levam a diversos riscos, desafios e impactos sociais, conforme mostrado nas tabelas a seguir. Baseamo-nos nos riscos e desafios identificados para modelos de base e, portanto, incluímos apenas os riscos, desafios e impactos sociais relacionados a agentes de IA que são novos ou que amplificam os da tabela de riscos dos modelos de base.

Observação: nas tabelas abaixo, o termo agente de IA é usado para descrever tanto o agente de IA quanto o sistema de IA agêntica.

Riscos

Grupo	Risco	Indicador de risco com justificativa
Alinhamento de valor	**Ações desalinhadas:** agentes de IA podem tomar ações que não estão alinhadas com valores humanos relevantes, considerações éticas, diretrizes e políticas. Ações desalinhadas podem ocorrer de diversas formas, como: • Aplicar metas aprendidas de forma inadequada a situações novas ou imprevistas. • Utilizar agentes de IA para propósitos/metas além de seu uso pretendido. • Selecionar recursos ou ferramentas de maneira enviesada. • Usar táticas enganosas para atingir o objetivo, desenvolvendo capacidade de • manipulação com base nas instruções fornecidas em um contexto específico. • Comprometer os valores do agente de IA para colaborar com outro agente ou ferramenta a fim de concluir a tarefa.	Amplificado Justificativa: • Autonomia dos agentes de IA para executar ações.

Justiça	**Ações discriminatórias:** os agentes de IA podem executar ações em que um grupo de pessoas é injustamente beneficiado em detrimento de outro, devido às decisões do modelo. Isso pode ser causado por vieses dos agentes de IA nas ações que impactam o mundo, nos recursos consultados e no processo de seleção desses recursos. Por exemplo, um agente de IA pode gerar código que contenha viés.	Amplificado
		
		Justificativa:
		• Autonomia dos agentes de IA para executar ações.
		  • Realização de ações que impactam o mundo
		  • Consulta a recursos enviesados
		  • Processo enviesado de seleção de recursos

	Viés de dados: ações específicas tomadas pelo agente de IA, como modificar um conjunto de dados ou banco de dados, podem introduzir viés no recurso que será utilizado por outras pessoas ou pelo próprio agente para executar ações.	Novo Justificativa: • Agentes de IA realizando ações que impactam o mundo. Aqui, o viés é devido a uma ação específica • Abertura
Confiança equivocada	**Confiança excessiva ou insuficiente:** a confiança — ou seja, a disposição para aceitar o comportamento de um agente de IA — depende do quanto o usuário confia nesse agente e para qual finalidade o utiliza. A confiança excessiva ocorre quando o usuário deposita confiança demais no agente de IA, aceitando seu comportamento mesmo quando provavelmente indesejado. A confiança insuficiente é o oposto, quando o usuário não confia no agente de IA, mesmo quando deveria confiar. O aumento da autonomia (para agir, selecionar e consultar recursos/ferramentas) dos agentes de IA, combinado com a possibilidade de opacidade e abertura, aumenta a variabilidade e a invisibilidade do comportamento dos agentes, dificultando a calibração da confiança e possivelmente contribuindo tanto para a superconfiança quanto para a subconfiança.	Amplificado Justificativa: • Ações de agentes de IA dificultam a avaliação de confiança

Grupo	Risco	Indicador de risco com justificativa
Ineficiência computacional	**Ações redundantes:** os agentes de IA podem executar ações que não são necessárias para atingir o objetivo. Essas ações podem desperdiçar recursos computacionais, reduzir a eficiência do agente para alcançar o objetivo e levar a resultados potencialmente prejudiciais. Em casos extremos, os agentes podem entrar em ciclos de execução repetitiva das mesmas ações sem progresso. Isso pode ocorrer por condições inesperadas no ambiente, falha do agente em refletir sobre suas ações, erros de raciocínio e planejamento ou falta de conhecimento sobre o problema. Isso pode impedir que o agente atinja o objetivo e exaurir recursos computacionais.	Novo Justificativa: • Agentes de IA podem executar ações
Robustez	**Ataque aos recursos externos dos agentes de IA:** agressores podem criar vulnerabilidades ou explorar vulnerabilidades existentes em recursos externos (ferramentas/bancos de dados/aplicações/serviços/outros agentes) dos quais o agente de IA depende para executar suas ações ou atingir seus objetivos. Recursos comprometidos podem impactar o desempenho do agente de IA de várias formas, como: • Manipular agentes de IA para perseguirem um objetivo diferente. Exemplo: alterar o objetivo do agente de IA para adicionar avaliações positivas a um produto de preferência do invasor, quando o objetivo original do usuário era adicionar perguntas sobre o produto. • Manipular agentes de IA para executarem ações indesejadas. Exemplo: enganar o agente de IA para que baixe um malware. • Capturar e retransmitir interações entre agentes de IA para agentes maliciosos. • Fazer com que agentes de IA compartilhem informações pessoais ou confidenciais.	Novo Justificativa: • Agentes de IA podem executar ações • Abertura e complexidades decorrentes da abertura • Acesso a mais recursos
	Uso não autorizado: se invasores conseguirem acesso ao agente de IA e seus componentes, poderão executar ações com diferentes níveis de dano, dependendo dos recursos do agente e das informações às quais ele tem acesso. Além disso, os invasores podem realizar ações que levem à degradação do sistema, como esgotamento de recursos disponíveis e redução de desempenho. Exemplos: • Utilização de informações pessoais armazenadas para imitar identidade ou se passar por alguém com intenção de enganar. • Manipulação do comportamento do agente de IA por meio de feedback ou corrupção da memória para alterar seu comportamento. • Manipulação da descrição do problema ou do objetivo para induzir o agente de IA a agir mal ou executar comandos nocivos.	Amplificado Justificativa: • Agentes de IA podem executar ações • Agentes de IA são mais capazes • Funcionalidade de personalização dos agentes de IA
	Explorar a discrepância de confiança: invasores podem iniciar ataques de injeção para ultrapassar o limite de confiança, que é um ponto ou linha conceitual onde muda o nível de confiança em um sistema, aplicação ou rede. Isso pode levar a limites de confiança divergentes (esperado vs. real), resultando em uso indevido de ferramentas, agência excessiva e escalonamento de privilégios. Além disso, a execução em segundo plano em ambientes com múltiplos agentes aumenta o risco de canais encobertos caso a validação de input/output seja fraca.	Amplificado Justificativa: • Abertura • Complexidade
	Alucinação em chamadas de função: agentes de IA podem cometer erros ao gerar chamadas de função (chamadas para ferramentas para executar ações). Essas chamadas podem resultar em ações incorretas, desnecessárias ou prejudiciais. Exemplos: geração de funções erradas ou parâmetros incorretos para as funções.	Novo Justificativa: • Os agentes de IA podem consultar recursos e ferramentas • Agentes de IA podem executar ações

Grupo	Risco	Indicador de risco com motivo
Privacidade e IP	**Compartilhamento de IP/PI/informações confidenciais com o usuário:** os agentes de IA com acesso irrestrito a recursos, bancos de dados ou ferramentas podem, potencialmente, armazenar e compartilhar informações pessoais, de propriedade intelectual ou confidenciais com os usuários do sistema ao executarem suas ações.	Amplificado Justificativa: • Natureza multicomponente com capacidade de executar ações
	Compartilhamento de IP/PI/informações confidenciais com ferramentas: agentes de IA com acesso irrestrito a recursos, bancos de dados ou ferramentas podem, potencialmente, armazenar e compartilhar informações pessoais, de propriedade intelectual ou confidenciais com outras ferramentas ou agentes ao executarem suas ações.	Novo Motivo: • Natureza multicomponente com a capacidade de executar ações
Explicabilidade e transparência	**Ações inexplicáveis e não rastreáveis:** explicações, rastreabilidade e informações de origem sobre as ações de agentes de IA podem ser difíceis, imprecisas ou impossíveis de obter.	Amplificado Justificativa: • Dificuldade em rastrear causa e efeito ou a influência de diferentes componentes, incluindo LLMs, nas ações finais
	Falta de transparência: a falta de transparência decorre da documentação insuficiente do design, desenvolvimento e processo de avaliação do agente de IA, além da ausência de informações sobre seu funcionamento interno e interação com outros agentes/ferramentas/recursos.	Amplificado Justificativa: • Confiança em outros documentos disponíveis para outras ferramentas/agentes

Desafios

Desafio	Indicador de desafio com motivo
Avaliação: desafio em avaliar o desempenho/precisão dos agentes de IA devido à complexidade do sistema e à sua natureza aberta.	Amplificado Justificativa: • Complexidade dos sistemas agênticos
Mitigação e manutenção: desafio em identificar onde algo está dando errado no sistema, como corrigir ou quais protocolos de manutenção devem ser aplicados.	Amplificado Justificativa: • Complexidade e abertura dos sistemas agênticos
Reprodutibilidade: desafio em reproduzir o comportamento ou a saída do agente devido à indisponibilidade ou mudanças nas ferramentas ou recursos usados para executar as ações.	Novo Justificativa: • Abertura
Responsabilidade: desafio na atribuição de responsabilidade por uma ação tomada por um sistema de IA agêntica.	Amplificado Justificativa: • Complexidade e abertura. Componentes podem ser de fornecedores diferentes
Conformidade: desafio em determinar a conformidade regulatória, já que os agentes de IA são complexos e pode não haver informações suficientes para entender se todo o sistema agente está em conformidade.	Amplificado Justificativa: • Abertura • Complexidade • Falta de transparência

impactos sociais

Impacto social	Indicador de impacto social
Impacto na dignidade humana: se os trabalhadores humanos perceberem que os agentes de IA são melhores do que eles em suas funções, isso pode causar uma diminuição em sua autoestima e bem-estar.	Amplificado
Impacto na autonomia humana: a natureza autônoma dos agentes de IA na execução de tarefas ou na tomada de decisões pode afetar a capacidade dos indivíduos de pensar criticamente, fazer escolhas e agir de forma independente.	Amplificado
Impacto nos empregos: a adoção generalizada de agentes de IA para executar tarefas complexas pode levar à automação em larga escala de funções e resultar em deslocamento de empregos.	Amplificado
Impacto no meio ambiente: a complexidade das tarefas e a possibilidade de os agentes de IA executarem ações redundantes podem levar a ineficiências computacionais e aumentar o impacto ambiental.	Amplificado

Mitigação e governança

À medida que a IA acelera a transformação dos negócios, a confiança se torna imperativa para navegar em um ambiente competitivo. O compromisso da IBM com a confiança está incorporado em nossos Princípios de confiança e transparência e nos Pilares de uma IA confiável, e este documento destaca a abordagem abrangente da IBM quanto à cultura, processos e ferramentas, além de como as equipes de pesquisa, produto e consultoria da IBM trabalham junto ao Conselho de Ética em IA para construir soluções de IA agêntica responsáveis.

Conselho de ética em IA

A IBM estabeleceu uma cultura que apoia o desenvolvimento, a implementação e o uso responsável da IA. No centro de nossa governança organizacional está o Conselho de Ética em IA, que é responsável pelo processo de governança e tomada de decisões sobre políticas e práticas de ética em IA. O Conselho de Ética em IA está em funcionamento há mais de cinco anos. Entre suas muitas atividades, o Conselho trabalha com pontos focais em cada unidade de negócios da IBM para avaliar casos de uso de IA e alinhá-los com os valores centrais da IBM.

Governança integrada

O Programa de Governança Integrada (IGP) é uma abordagem unificada para responsabilidade e conformidade. Ao criar uma visão abrangente de ponta a ponta dos dados e modelos criados e utilizados, o IGP escalona fluxos de trabalho de governança em torno de dados, privacidade e IA sem interromper os processos de inovação e negócios. O IGP permitiu à IBM implementar padrões internos de dados uniformes, promovendo transparência dos dados e desenvolvimento de IA confiável, ao mesmo tempo em que possibilita inovação em escala.

Produtos e ofertas

O IBM watsonx.governance® permite que as organizações conduzam uma IA responsável, transparente e explicável. Ele oferece recursos de governança completos ao longo de todo o ciclo de vida da IA, incluindo IA agêntica, desde a solicitação do caso de uso até a implementação, incluindo avaliação de riscos iniciais e avaliação de riscos para ajudar a identificar riscos logo no início do processo. Possui recursos de geração aumentada de recuperação (RAG) e métricas de avaliação de IA agêntica, como fidelidade, relevância do contexto e similaridade de resposta, para ajudar a confirmar se os agentes de IA estão agindo adequadamente. O watsonx.governance da IBM terá métricas adicionais projetadas para monitorar e melhorar o desempenho dos agentes. Também contará com recursos para detectar alucinações em chamadas de ferramentas, gerenciar riscos, auxiliar na conformidade regulatória e capturar metadados e outros detalhes sobre os agentes de IA em uma ficha técnica.

O IBM watsonx.ai simplifica, unifica e otimiza o gerenciamento do ciclo de vida dos agentes (conhecido como AgentOps), fornecendo transparência, rastreabilidade e flexibilidade para descobrir, gerenciar, monitorar e otimizar agentes de IA.

O IBM watsonx Orchestrate coloca a IA em ação ao ajudar os usuários a criar, implementar e gerenciar agentes de IA poderosos que automatizam tarefas com IA generativa. O Orchestrate fornecerá transparência sobre o raciocínio dos agentes de IA na seleção de ferramentas e/ou colaboração com outros agentes, permitindo que o usuário entenda como uma tarefa ou fluxo de trabalho foi concluído.

O IBM® Guardium AI Security ajuda os clientes a monitorar continuamente os controles de seus modelos de IA generativa em produção e permite uma implementação segura e responsável.

A oferta de estratégia e governança de IA da IBM Consulting permite que as empresas aproveitem o potencial transformador de uma IA selecionada de forma responsável — desde a IA convencional até a IA agêntica — ancorada em seus dados empresariais, para avançar e remodelar sua estratégia de negócios. Aconselhamos nossos clientes a “aplicar a IA certa” para otimizar o ROI, e a “lidar com a IA corretamente”, garantindo responsabilidade sobre os modelos de IA que desenvolvem e adquirem, para que possam escalar com responsabilidade enquanto limitam os riscos associados ao investimento. Esse framework abrangente de requisitos permite que os clientes enfrentem os desafios em evolução da IA agêntica, descubram novas oportunidades, acelerem a inovação e aprimorem a tomada de decisão dentro da organização. A seguir, apresentamos as práticas e recomendações integradas à oferta para ajudar as empresas a reduzir os riscos associados à IA agêntica:

- Ativamos observabilidade para entender quais ações o agente realizou com quais inputs, a fim de determinar a atribuição para as saídas correspondentes.
- Examinamos todo o rastreamento para determinar se o agente de IA avançou em direção ao objetivo geral com cada ação tomada, para depurar o fluxo e identificar gargalos no design do sistema que frequentemente geram ações redundantes ou desnecessárias.
- Construímos ontologias para melhorar a precisão, confiabilidade, além de fornecer linhagem e proveniência dos dados.
- Realizamos testes funcionais que envolvem testar rigorosamente cada componente do sistema agente de forma isolada (isto é, ferramentas, agentes), interações entre os componentes, a capacidade de selecionar ferramentas apropriadas com os parâmetros e valores corretos, e barreiras de proteção para mitigar vulnerabilidades.
- Configuramos agentes de IA para melhorar a eficiência computacional detectando quando os agentes entram em possíveis cenários de loop infinito, acompanhando, por tarefa, o tempo decorrido, total de tokens consumidos ou número de iterações sem sucesso.
- Criamos agentes de IA hiperfocados e fornecemos a eles apenas as ferramentas adequadas e necessárias para concluir suas tarefas, garantindo que a execução seja sempre feita no contexto de autorização do usuário que acessa.
- Definimos barreiras de proteção no nível do modelo para detectar e mitigar conteúdo HAP, jailbreaks, tentativas de injeção de prompt, divulgação não autorizada de informações sensíveis e alucinações.

Modelos

Os modelos Granite Guardian são um conjunto robusto de salvaguardas projetadas para detectar riscos tanto em prompts quanto em respostas. Em casos de uso com RAG, os modelos Guardian avaliam a relevância do contexto, fundamentação e relevância da resposta. Eles também contam com detectores de alucinações em chamadas de função dentro de fluxos de trabalho agênticos. Isso inclui a avaliação da validade das chamadas de função e a detecção de informações fabricadas, especialmente durante a tradução de consultas.

Supervisão humana e validação por humanos

A supervisão e revisão humanas podem ajudar a identificar riscos e corrigir erros. A validação e o feedback humanos ajudam a garantir que as ações realizadas pelos agentes de IA sejam precisas, relevantes e alinhadas. Como parte do processo de avaliação ética de casos de uso de IA da IBM, a questão da supervisão humana é considerada, e a supervisão apropriada é implementada. O ponto de vista da IBM sobre o reforço da inteligência humana apresenta exemplos de casos de uso, indicadores-chave de desempenho e melhores práticas para aprimorar a inteligência humana com IA e capacitar indivíduos a navegar em um ambiente de negócios competitivo em parceria com a IA.

Educação

A IBM oferece educação sobre ética e governança de IA para equipes, clientes e comunidades. O IBM SkillsBuild, canal IBM Technology no YouTube, os cursos watsonx para desenvolvedores da IBM e o IBM AI Academy são alguns dos recursos por meio dos quais a IBM promove a educação sobre ética e governança de IA.

A seguir estão destaques de pesquisas e ferramentas de ponta da IBM para ajudar usuários ao longo do ciclo de vida de agentes de IA e desenvolver soluções responsáveis de IA agêntica.

Técnicas e métodos

- Técnicas como forçar os humanos a tomarem mais tempo para pensar (estratégia de desancoragem baseada no tempo) ajudam a alcançar uma colaboração ideal entre humanos e IA e reduzir o viés de ancoragem, quando os humanos confiam cegamente na decisão da IA. Outra técnica baseia-se no framework colaborativo IA-humano baseado em valores, que introduz fricção ao orientar humanos com recomendações de decisão, quando necessário, em interações onde o humano é o tomador de decisão final.
- Métodos como Explicação em múltiplos níveis para modelos de linguagem generativa e Explicações contrastivas para modelos de linguagem de grande escala auxiliam na explicabilidade e atribuição de origem.
- Métodos como colaboração adversarial, em que a IA examina a base da decisão humana em vez de oferecer recomendações alternativas, ajudam a enfrentar o impacto da automação e da redução da agência sobre a dignidade humana.
- Attack Atlas, uma taxonomia intuitiva e organizada de vetores de ataque de input de uma única rodada, oferece à comunidade um ponto de partida unificado no campo em rápido crescimento da segurança em IA generativa e red teaming.

Ferramentas e benchmarks

- Abordagens de ajuste de modelo, como as empregadas pelo IBM Alignment Studio, ajudam a mitigar os riscos de alinhamento de valores.
- O kit de ferramentas de código aberto da IBM AI Fairness 360 ajuda a mitigar riscos relacionados à equidade.
- O ITBench, um conjunto de benchmarks, oferece aos profissionais de IA uma forma de medir a eficácia dos agentes que estão desenvolvendo para resolver problemas reais e como seus agentes se comparam a outros em tarefas do dia a dia dos negócios.
- Carbon for AI, um sistema de design de código aberto da IBM, utiliza um ícone de IA interativo para promover a explicabilidade e uma compreensão mais clara dos recursos da IA.

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

Produzido nos
Estados Unidos da América
Março de 2025