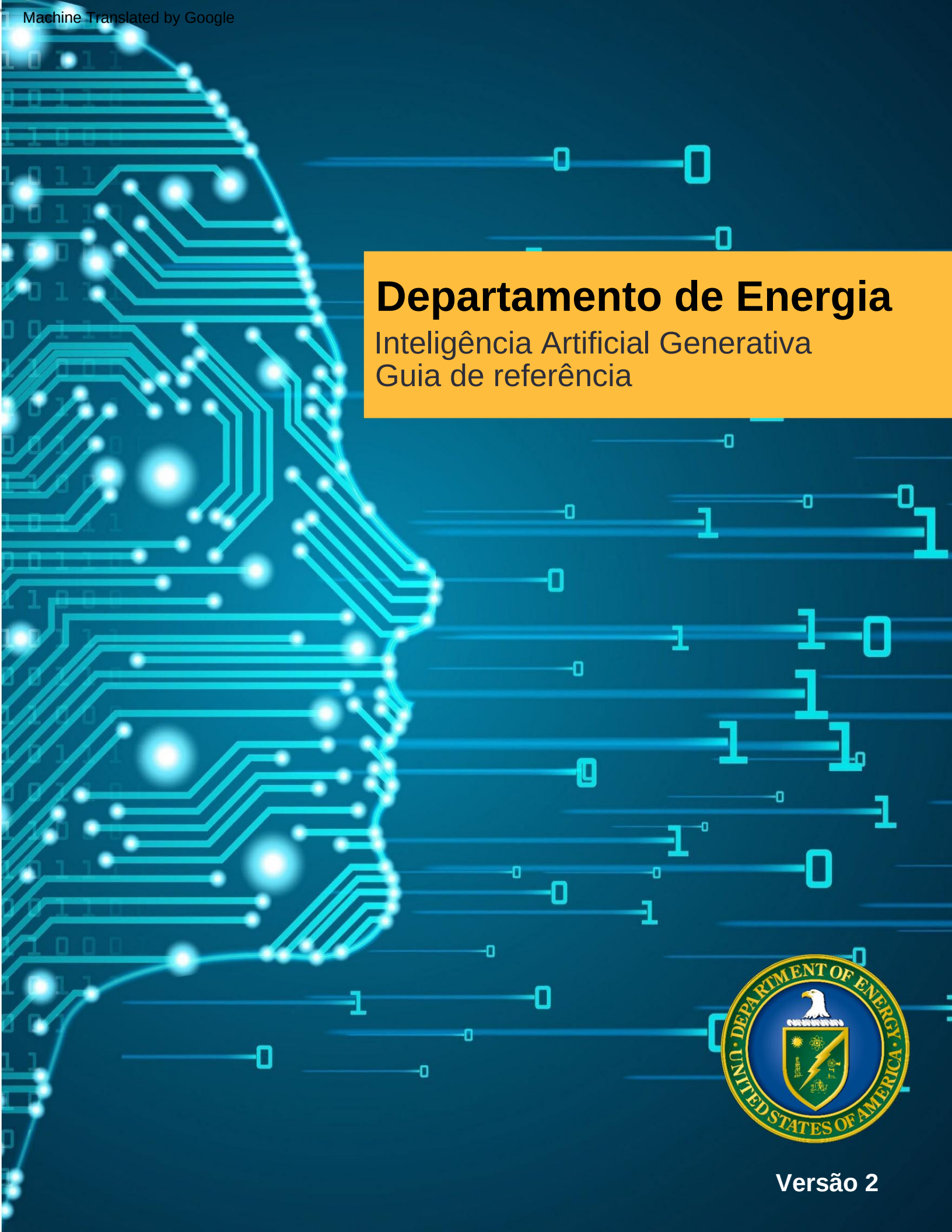
Departamento de Energia

Inteligência Artificial Generativa
Guia de referência



Versão 2

Registro de Alterações

Data da versão		Autor/Proprietário	Descrição da mudança
Versão 1.0	22 de setembro de 2023	Gabinete do Chefe Oficial de Informação	Guia de referência de IA generativa do DOE v1 Para divulgação interna
Versão 2.0	26 de abril de 2024	Gabinete do Chefe Oficial de Informação	Guia de referência de IA generativa do DOE v2 Para divulgação pública

Índice

U.S. DEPARTMENT OF ENERGY

1. Visão geral do documento

ÿ Este é um guia de referência para o uso de IA generativa e não deve ser interpretado como uma política. Como tal, não prescreve ações específicas.

ÿ Este documento foi desenvolvido com um público geral em mente e não inclui atualmente considerações direcionadas para funções especializadas, incluindo equipe de Pesquisa e Desenvolvimento (P&D) e Gestão e Operação (M&O). Considerações mais profundas para essas funções podem ser abordadas em uma próxima iteração deste documento.

ÿ A IA generativa (GenAI) é uma ferramenta incrivelmente poderosa que tem enorme potencial para permitir o progresso científico, aumentar a produtividade da força de trabalho do Departamento de Energia (DOE) e impulsionar a missão de inovação do DOE com tecnologias emergentes.

ÿ O GenAI é mais bem utilizado para fornecer um primeiro rascunho ou para ajudar a encontrar opções ou alternativas, em vez de ser usado para produzir um resultado final preciso e imparcial.

ÿ De acordo com a Ordem Executiva 14110 sobre o *Desenvolvimento e Uso Seguro, Protegido e Confiável de Inteligência Artificial,* as agências federais são desencorajadas a impor proibições ou bloqueios gerais amplos ao uso de IA generativa por agências. O DOE está no processo de considerar quais serviços GenAI serão permitidos para uso com base em avaliações de risco abrangentes. À medida que as decisões sobre os serviços forem tomadas, diretrizes específicas para uso serão estabelecidas.

ÿ Todas as regras existentes do DOE referentes ao gerenciamento e uso de dados devem ser seguidas. Entre em contato com o OCIO ou com o Diretor de Privacidade do DOE para perguntas. Perguntas legais devem ser direcionadas ao Conselheiro Geral Assistente do DOE para Transferência de Tecnologia e Propriedade Intelectual ou a um consultor jurídico experiente do contratante.

ÿ Continue a usar o bom senso e a seguir as regras existentes relativas a dados e informações gerenciamento ao usar o GenAI.

ÿ Ter um humano no circuito para revisar os resultados quanto à precisão, considerações éticas, qualidade e verificar possível viés.

ÿ Consulte orientações federais específicas (Seção 4).

ÿ Consulte as principais aplicações e casos de uso para entender exemplos de como o GenAI pode ser aplicado para impulsionar valor e inovação no DOE (Seção 6).

ÿ Sua função na organização (ou seja, usuário geral, cientista de dados, liderança) é um fator-chave considerações e melhores práticas que são mais relevantes para você (Seção 7.2).

ÿ Tenha em mente as principais considerações e as melhores práticas para gerenciar adequadamente os riscos associados a IA e GenAI (Seção 8).

ÿ Consulte a Lista de verificação de melhores práticas para orientar seu uso do GenAI (Seção 8.11).

ÿ Qualquer referência a um modelo ou produto GenAI específico neste documento não deve ser interpretada como uma endosso do modelo ou de qualquer um dos seus resultados potenciais.

ÿ À medida que a GenAI continua a evoluir, o DOE terá que permanecer ágil e se ajustar ao cenário em constante mudança de oportunidades, riscos e melhores práticas. Esta orientação será atualizada regularmente para refletir o pensamento mais atual.

2. Resumo Executivo

O Guia de Referência de Inteligência Artificial Generativa do Departamento de Energia (DOE) versão 21 está sendo emitido como uma referência sobre IA generativa (GenAI), uma tecnologia de IA relativamente mais nova que pode produzir vários tipos de conteúdo, para todo o complexo do DOE, incluindo funcionários federais e contratados em laboratórios e locais do DOE. Principais partes interessadas e especialistas no assunto (SMEs) de toda a organização do DOE estabeleceram uma equipe de tigres para colaborar no desenvolvimento deste documento. O esforço coordenado forneceu uma variedade de perspectivas de várias funções e papéis do DOE que são entrelaçados por toda parte. A colaboração e o envolvimento contínuos de uma variedade de partes interessadas beneficiarão futuras iterações deste documento e impulsionarão a inovação da IA no DOE. Este documento não é uma política ou diretiva, mas sim um guia de referência para ajudar as partes interessadas de todo o DOE a entender como usar o GenAI de forma responsável. Este documento e as orientações contidas nele serão atualizados regularmente à medida que a tecnologia GenAI e o ambiente regulatório que a cerca continuarem a evoluir. À luz da complexidade do GenAI e do ritmo em que as pesquisas e os avanços comerciais estão sendo feitos, alavancar a expertise de pesquisadores e SMEs será vital. Este guia não substitui aconselhamento jurídico; portanto, quaisquer questões jurídicas relacionadas ao uso do GenAI devem ser direcionadas ao consultor jurídico competente do DOE ou do contratante.

> *"As pessoas às vezes perguntam: 'A IA é nossa amiga ou nossa inimiga?' E minha resposta para isso é: acho que a IA é nossa amiga, mas, assim como qualquer bom relacionamento, há limites."2*
>
> - Gardy Rosius, vice-CIO do DOE

O GenAI promete promover a missão do departamento, mas também apresenta riscos. Ao empregar o GenAI, é preciso estar ciente das capacidades e limitações da tecnologia e deve-se ter em mente que o usuário, não a tecnologia GenAI, continua sendo responsável por quaisquer ações ou resultados resultantes do uso das tecnologias GenAI.

Portanto, os usuários não devem confiar nos sistemas GenAI para tomar decisões; em vez disso, eles devem usar os sistemas para informá-los.

Este documento compreende informações úteis que podem ser usadas para disseminar a conscientização em todo o DOE sobre o uso responsável do GenAI. Os tópicos incluem o histórico do GenAI, um resumo das leis e mandatos existentes referentes ao GenAI (no momento da publicação), tópicos fundamentais sobre o uso responsável do GenAI (incluindo funções organizacionais, dados e modelos de serviço), casos de uso em potencial e os riscos e melhores práticas mais proeminentes em torno desta tecnologia emergente.

De cientistas de dados a liderança e usuários em geral, todos no DOE têm um papel a desempenhar no uso responsável da tecnologia GenAI. Depois de ler este documento, o leitor deve ter uma consciência recém-descoberta ou aumentada de seu papel no uso responsável da GenAI, bem como conhecimento fundamental das soluções GenAI, as principais considerações e riscos que devem ser contabilizados e as melhores práticas atuais para mitigar riscos e usar a tecnologia GenAI de forma responsável.

3. Objetivo e escopo

O objetivo deste documento é fornecer uma compreensão dos principais benefícios, considerações, riscos e melhores práticas associadas ao GenAI no contexto do DOE. Este documento pretende servir como uma referência valiosa sobre o GenAI para todos os grupos dentro do ambiente do DOE, oferecendo uma visão geral dos riscos, considerações, responsabilidades e recomendações específicas que estão associadas a vários
funções organizacionais.

O escopo deste documento é que ele serve como a segunda versão de um guia de referência destacando riscos específicos do GenAI e melhores práticas. Este documento é um verdadeiro guia de referência, não indicativo de uma política ou diretiva. As melhores práticas recomendadas neste documento não substituem nenhuma lei, regulamento,

ou políticas existentes do DOE. Como tal, este documento inclui uma discussão de informações básicas, conceitos e definições principais e oportunidades para aplicar o GenAI, bem como uma discussão de considerações principais, riscos, melhores práticas e recomendações. Este documento *não* inclui nenhuma ação prescritiva e obrigatória, pois estas serão capturadas em políticas existentes e futuras. Além disso, este documento tem como objetivo complementar, mas não substituir, as regulamentações existentes em torno do GenAI.

4. Diretrizes e referências federais

O GenAI está evoluindo rapidamente: sua tecnologia subjacente continua a avançar, a variedade de ferramentas GenAI disponíveis no mercado continua a crescer e o GenAI está se tornando cada vez mais acessível ao público. À medida que essa evolução acelera, cresce a necessidade de conscientização sobre os impactos potenciais da GenAI, bem como de identificação e mitigação dos riscos associados.

Várias publicações federais sobre IA e GenAI foram emitidas nos últimos anos. Esses documentos são o primeiro ponto de referência para este documento, fornecendo proteções para como o GenAI pode ser usado no governo federal. Notavelmente, a Ordem Executiva 14110 sobre o *Desenvolvimento e Uso Seguro, Protegido e Confiável de Inteligência Artificial* foi publicada recentemente em 30 de outubro de 2023. A EO 14110 contém uma variedade de diretivas que se aplicam ao Departamento de Energia, incluindo ações que o DOE é obrigado a liderar ou é obrigado a colaborar com outras agências para entregar. O Diretor do Escritório de Gestão e Orçamento emitiu um Memorando para os Chefes de Departamentos Executivos e Agências sobre o Avanço da Governança, Inovação e Gestão de Riscos para o Uso de Inteligência Artificial por Agências em março de 2024, que também deve servir como um documento de referência importante.

Este guia de referência não substitui a lei ou política existente e não pretende entrar em conflito com nenhuma legislação pendente relevante. Como acontece com todas as políticas, incluindo aquelas não discutidas neste documento, os membros da equipe devem revisar e continuar a aderir às políticas, procedimentos e guias do DOE para garantir a conformidade com os requisitos de informações do laboratório/DOE. Os funcionários também devem continuar a seguir os requisitos existentes, como aqueles relacionados à qualidade, segurança da informação e integridade da pesquisa. Os membros da equipe devem trabalhar com PMEs e organizações de conformidade de laboratório/DOE apropriadas, como o Office of General Counsel (GC), o Office of Export Control, o Classification Office, o Office of Meio Ambiente, Saúde, Segurança e Proteção (EHSS) e outros conforme apropriado. O guia de referência será atualizado conforme novas políticas e diretrizes forem emitidas. Resumos e detalhes selecionados das referências abaixo podem ser encontrados no Apêndice E no final deste documento.

Os recursos e referências federais relevantes existentes incluem:

1. Memorando do Gabinete de Gestão e Orçamento para os Chefes dos Departamentos Executivos e Agências, promovendo a governança, a inovação e a gestão de riscos para o uso de inteligência artificial por agências Inteligência, março de 2024
2. Ordem Executiva 14110 sobre Desenvolvimento e Uso Seguro, Protegido e Confiável de Inteligência Artificial, outubro de 2023
3. Inteligência Artificial Generativa e Privacidade de Dados: Uma Introdução, Congressional Research Service (CRS), Maio de 2023
4. Inteligência Artificial Generativa e Lei de Direitos Autorais, Congressional Research Service (CRS), maio 2023
5. Relatório do Ano 1 do Comitê Consultivo Nacional de Inteligência Artificial (NAIAC), maio de 2023
6. Estrutura de Gestão de Riscos de IA, Instituto Nacional de Padrões e Tecnologia (NIST), janeiro 2023
7. Lei de promoção da IA americana, dezembro de 2023

8. Treinamento de IA para a Lei de Aquisição de Força de Trabalho, outubro de 2022

9. Projeto para uma Declaração de Direitos da IA, Escritório de Política Científica e Tecnológica (OSTP), outubro de 2022

10. Secure Software Development Framework (SSDF V.1,1), NIST, fevereiro de 2022

11. Estrutura de responsabilidade da IA para agências federais, GAO, junho de 2021

12. Lei da Iniciativa Nacional de IA, janeiro de 2021

13. Ordem Executiva 13960 sobre a promoção do uso de IA confiável no governo federal, Dezembro de 2020

14. Lei de IA no Governo, setembro de 2020 15. Ordem Executiva 13859 sobre a manutenção da liderança americana em IA, fevereiro de 2019

16. Lei de Autorização de Defesa Nacional John S. McCain, Seção 1051 para o Ano Fiscal de 2019

17. Lei do Governo Eletrônico de 2002

Consulte Congress.gov para visualizar o status da legislação de IA proposta e pendente.

5. Antecedentes da Inteligência Artificial Generativa

5.1 IA, IA generativa e GPT

A inteligência artificial avançou tremendamente desde que foi introduzida pela primeira vez na década de 1950. Seu crescimento superou dois patamares de avanço que ocorreram quando a visão para a aplicação da IA era mais ampla do que a capacidade funcional na época (ou seja, não havia poder de computação ou dados suficientes e nenhum algoritmo suficientemente avançado para operacionalizar a visão). Nos últimos anos, a IA ganhou cada vez mais atenção pública, tornando-se um tópico quente na tecnologia, bem como nos Estados Unidos e no mundo.
sociedade.

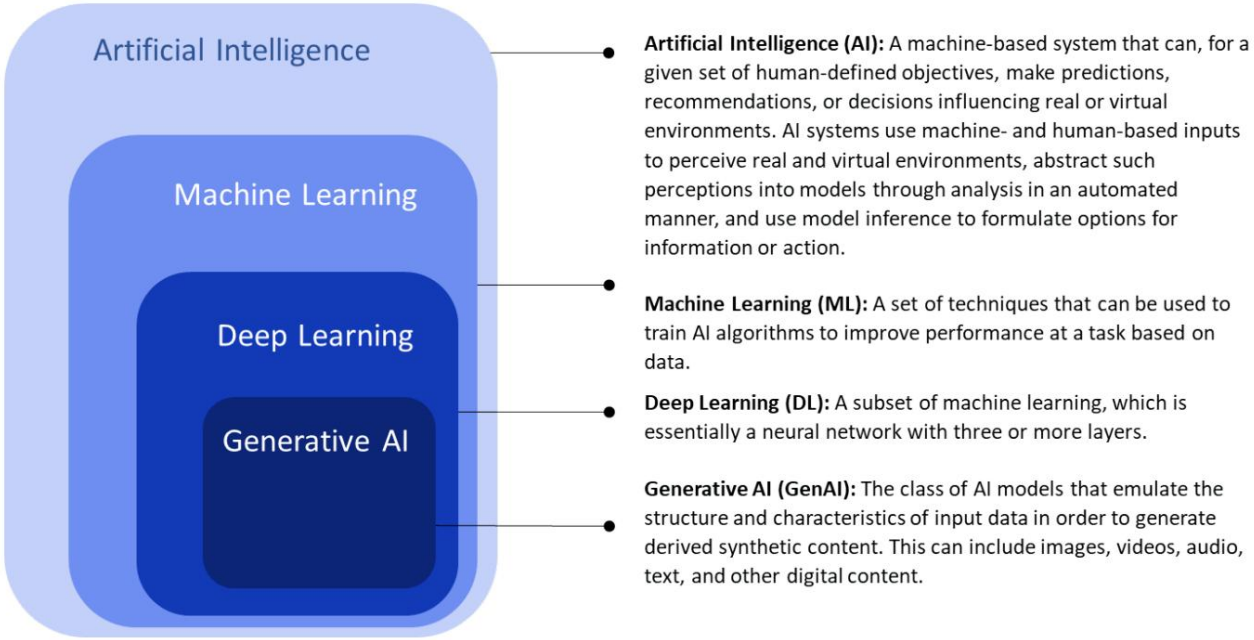


Figura 1: Definições ilustrativas de inteligência artificial, machine learning, deep learning e IA generativa. Fontes de definição: Inteligência Artificial,3 Machine Learning,4 Generative AI,5 e Deep Learning6

Assim como a IA, **a GenAI** não é nova, mas vem ganhando força desde a introdução de redes adversárias generativas (GANs), um tipo de algoritmo de aprendizado de máquina, em 2014. Esse desenvolvimento permitiu a criação de modelos generativos de imagem. Dois avanços recentes adicionais, **transformadores** e **modelos de linguagem grande (LLMs)** aceleraram ainda mais a evolução e adoção da GenAI.

Os transformadores são um modelo de aprendizado profundo que adota o mecanismo de autoatenção, ponderando diferencialmente a significância de cada parte dos dados de entrada.7 Em essência, eles são uma técnica que busca ajudar os modelos de IA a determinar no que prestar atenção.	**Large language models (LLMs)** usam aprendizado auto-supervisionado para aprender com grandes quantidades de dados de texto não estruturados e não rotulados. Esses modelos são treinados em grandes corpos de dados, permitindo que um modelo seja usado para uso múltiplo casos.

O surgimento do transformador em 2017, bem como o progresso feito com convoluções e recorrências para desempenho e velocidade de treinamento, levou ao **transformador pré-treinado generativo (GPT)** evoluindo para os LLMs de hoje. O GPT, o tipo de IA que tem estado no centro da atividade mais visível nos últimos anos, é baseado em redes neurais, que são um tipo de modelo de aprendizado de máquina (ML) construído para imitar as redes neurais biológicas que compõem os cérebros de humanos e animais.

GPT é uma família de LLMs construída em **rede neural profunda (DNN)** arquitetura que foi ajustada usando técnicas **de processamento de linguagem natural (PLN)** e **aprendizado por reforço de feedback humano (RLHF) ,** conforme ilustrado na Figura 2.
ChatGPT é o modelo de IA de última geração voltado para o consumidor, construído no GPT. Ele pode responder a perguntas solicitadas pelo usuário, gerar histórias, resumir texto como livros ou artigos e pesquisar texto com base em consultas conceituais. Observe que o ChatGPT está atualmente disponível no DOE para uso mediante solicitação com base na necessidade da missão. Guardrails adicionais podem ser desenvolvidos e implementados no futuro, conforme apropriado.

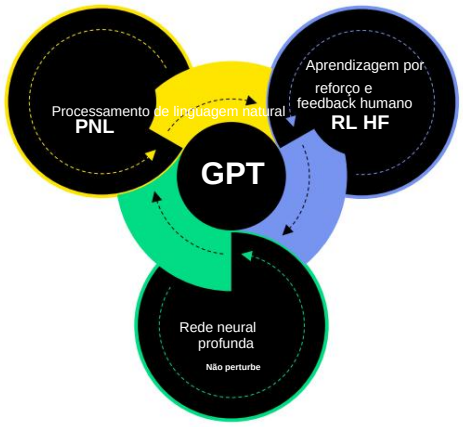


Figura 2: Representação ilustrativa do GPT

Os modelos de fundação, conforme denominados pelos pesquisadores da Universidade de Stanford, são treinados em grandes quantidades de dados não rotulados usando um algoritmo transformador que pode ser ajustado para uma ampla gama de tarefas posteriores. Para especializar ainda mais os modelos, os cientistas de dados podem treinar ou ajustar independentemente um modelo de fundação para construir **modelos específicos de tarefas,** que são modelos projetados para serem eficazes em tarefas específicas. A Figura 3 mostra o relacionamento de alto nível entre a fundação modelos e modelos específicos de tarefas. Definições adicionais relacionadas à IA podem ser encontradas no Apêndice K: Glossário.

Conforme ilustrado na Figura 3, usando modelos de fundação como ponto de partida e incluindo técnicas como ajuste fino supervisionado, ajuste de instruções e RLHF, modelos específicos de tarefas que se encaixam na situação em questão são construídos. A situação em questão pode incluir especificidades do caso de negócios, modalidades (por exemplo, texto, imagem/vídeo, fala, codificação automática, etc.), arquitetura de solução, dados específicos do caso de uso e uso pretendido. Desde que a OpenAI lançou o ChatGPT

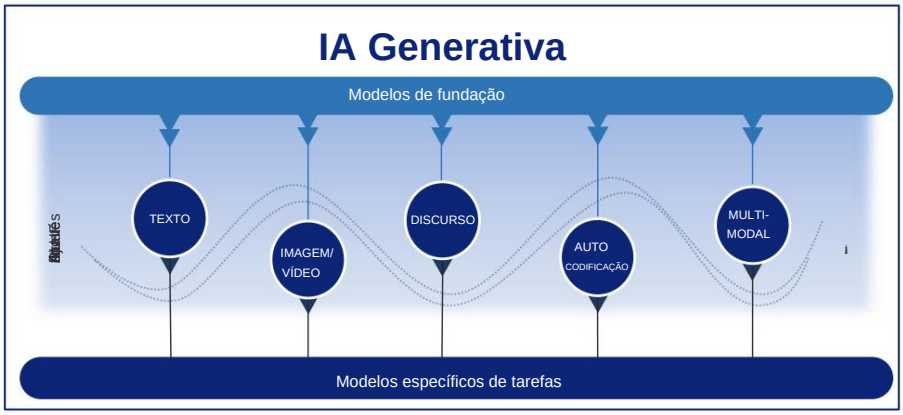


Figura 3: GenAI: modelos de base vs. modelos específicos de tarefas

em novembro de 2022, novos modelos GenAI construídos para serem específicos de tarefas — especializando-se em diferentes indústrias, subindústrias ou tipos de aplicações funcionais — têm entrado rapidamente no mercado e são geralmente genéricos ou construídos no local em um ambiente localizado. A facilidade de uso e o fácil acesso pelo público em geral pela internet ajudaram a tornar os modelos GenAI cada vez mais populares.

A maioria dos modelos é unimodal, o que significa que eles focam em uma única forma de informação, como texto, fala ou código de computador. Modelos multimodais podem aprender de múltiplas formas de entrada e produzir múltiplas formas de saída. Consulte a Tabela 1 abaixo para uma lista das várias modalidades e uma lista de amostra de aplicativos e modelos específicos de tarefas (ou personalizados) atualmente disponíveis no mercado. Observe que a Tabela 1 não distingue entre modelos de base e específicos de tarefas.

Categoria	Modalidade	Descrição	Exemplos de aplicações	Exemplos de modelos específicos de tarefas
Texto	**Unimodal**	Geração de texto semelhante ao humano a partir de prompts de texto	ChatGPT, Bardo, Claude 2, Bing	Jasper, cópia.ai, NukeLM8
Imagem/vídeo	**unimodal**	Geração de várias imagens e vídeos com base em prompts de texto	DALL-E 2, Midjourney, Difusão Estável (Automático1111), Estabilidade.ai	Midjourney, Craiyon, Estável LM 2 1.6B
Fala	**Unimodal**	Geração de fala sintetizada a partir de prompts de texto, reconhecimento de fala	Conteúdo do trovão, Voz limpa	Síntese de voz, podcast.ai, Speechmatics
Codificação automática	**unimodal**	Geração de código (por exemplo, Python, Java, JavaScript) a partir de prompts de texto	Copiloto do GitHub, Amazon Sussurrador de Código, Codebots, códigos OpenAI, ChatGPT, Bard	Copiloto do GitHub, Tabnine, Cograma
Multimodal	Geração multimodal de várias saídas onde o modelo aprende de uma variedade de fontes, incluindo texto, imagens e áudio		Gato, Mural do Google, GPT-4, GPT-5	Serviço Azure Open AI, Google Vertex IA, AWS Soluções, IBM Garage

Tabela 1: Modalidades GenAI

5.2 Tendências

O GenAI e suas técnicas subjacentes estão evoluindo e avançando rapidamente, e a adoção do GenAI está explodindo em um ritmo semelhante. Grandes avanços já foram feitos com a tecnologia GenAI desde sua introdução. Por exemplo, a OpenAI lançou o GPT-4 em março de 2023 e, em julho de 2023, uma marca registrada foi registrada para o GPT-5, o que sugeriu uma variedade de novos recursos potenciais para a próxima iteração do modelo de linguagem. A lista inclui recursos que expandem o ChatGPT além do GenAI de texto para texto e para o espaço multimodal, incluindo produção artificial de fala e texto humanos, conversão de áudio para texto, reconhecimento de voz e fala e desenvolvimento e implementação de redes neurais artificiais.9 Observe que muitas dessas funcionalidades, como reconhecimento de fala, são anteriores ao surgimento do GenAI, mas agora podem ser aprimoradas por meio do GenAI. As soluções GPT devem continuar a avançar em um ritmo agressivo.

A GenAI está prevista para ser um elemento importante no mundo profissional nos próximos anos e para atingir desempenho de nível humano mais cedo do que o previsto anteriormente. A Gartner prevê que…

- Até 2026, 75% das empresas usarão IA generativa para criar dados sintéticos de clientes, acima dos menos de 5% em 2023.
- Até 2027, mais de 50% da GenAI os modelos que as empresas usam serão específicos de domínio — específicos para um setor ou função empresarial — acima de aproximadamente 1% em 2023.
- Até 2027, mais da metade da seleção de ativos de desenvolvimento de mercados de tecnologia será realizada por orquestração de IA generativa.

Figura 4: Fonte: Gartner®, “Predicts 2024: The Future of Tecnologias de IA generativas”, Arun Chandrasekaran, Anthony Mullen, Lizzy Foo Kune, Nicole Greene, Jim Hare, Leinar Ramos, Anushree Verma, 28 de fevereiro de 202410

Da mesma forma, entre março de 2023, quando a solução GenAI Claude da Anthropic foi lançada no mercado, e maio de 2023, grandes avanços foram feitos na velocidade de processamento da solução.

O GenAI já está transformando rapidamente áreas como marketing e mídia, enquanto em outras áreas, ele ainda está em um estado emergente. A lista de casos de uso em potencial (explorados mais detalhadamente na Seção 6: Oportunidades para aplicar o GenAI) continua a crescer à medida que o GenAI continua progredindo em suas habilidades para gerar múltiplas formas de mídia, incluindo texto, imagem, vídeo, fala, música e código de programação.

Embora o GenAI já tenha ganhado uma enorme quantidade de tração de uma multidão de organizações e em uma miríade de aspectos da sociedade, ele ainda está realmente em sua infância. Espere desenvolvimentos rápidos com os recursos do GenAI e com a proliferação de seus potenciais casos de uso e aplicações para continuar. À medida que o GenAI continua a evoluir, o mercado e as organizações que o adotam terão que permanecer ágeis e se ajustar ao cenário em constante mudança de oportunidades, regulamentações, riscos e melhores práticas.

A rápida evolução do GenAI traz muitos benefícios potenciais, mas também muitos riscos e efeitos desconhecidos. Embora uma variedade de riscos relacionados ao GenAI já tenham surgido, espere que alguns riscos se tornem mais pronunciados e novos riscos apareçam conforme a adoção do GenAI acelera. Estabelecer estratégias de gerenciamento de risco, documentar e compartilhar as melhores práticas e incentivar a conscientização em toda a organização sobre os riscos e recomendações associados ao GenAI serão etapas críticas para adotar o GenAI com sucesso e perceber os muitos benefícios que ele pode oferecer.

5.3 Proposta de Valor

Os casos de uso e as aplicações potenciais do GenAI estão crescendo rapidamente. Simplificando, o valor do GenAI é preencher a função de um “copiloto” automatizado para criar materiais em várias formas de mídia, incluindo texto, imagem, vídeo e código de programação. Dentro do DOE, isso significa que o GenAI pode ser capaz de aumentar o trabalho produzido por humanos com velocidade. Uma vez adotado, o GenAI pode mudar as funções humanas existentes dentro da organização sem necessariamente substituí-las. A Seção 6 da Ordem Executiva 14110 inclui uma variedade de mandatos sobre a exploração dos efeitos da IA nos direitos dos trabalhadores e na estabilidade econômica. Informações adicionais podem ficar disponíveis à medida que esse relatório for concluído.

Quando hipoteticamente usado como um copiloto para funcionários do DOE, o GenAI tem o potencial de ajudar os funcionários com tarefas do dia a dia, incluindo (mas não se limitando a) encontrar informações mais rapidamente com sua funcionalidade de pesquisa, gerar resumos de reuniões e documentos longos e redigir e-mails e outras correspondências. Esses exemplos simples são áreas em que a tecnologia GenAI já é proficiente em gerenciar certas tarefas muito rapidamente e em escala. O GenAI pode ser capaz de produzir esboços de pesquisa ou conteúdo e pontos de partida para o conteúdo para permitir que os funcionários do DOE tenham mais tempo para se concentrar no refinamento e desenvolvimento do produto. O futuro local de trabalho provavelmente incluirá um relacionamento simbiótico entre funcionários humanos e o GenAI. À medida que as tecnologias de IA se integram às ferramentas de trabalho do dia a dia (um exemplo pode eventualmente incluir o Office365) e, portanto, são menos visíveis para o usuário, esse relacionamento pode mudar ou exigir exploração adicional de risco e permissões de uso.

As soluções GenAI podem executar tarefas de rotina muito mais rápido do que os humanos (embora isso introduza riscos em relação à precisão, confiabilidade e “alucinações”, que são discutidos na Seção 8: Principais Considerações e Melhores Práticas). A GenAI pode criar mais tempo e espaço para os funcionários do DOE agregarem valor ao seu trabalho, capacitando-os a otimizar seu tempo durante sua semana de trabalho.

Espera-se que o GenAI forneça capacidades que permitirão ao DOE inovar mais rapidamente. Por exemplo, o GenAI pode usar grandes conjuntos de dados e conteúdo relativamente inexplorados para derivar insights acionáveis que podem ajudar a impulsionar o valor do negócio.

Há quatro funções primárias da capacidade de texto para texto do GenAI. Qualquer uma dessas funções pode ser usada sozinha ou pode ser “agrupada” para uma solução. Entender essas quatro funções pode ajudar a explicar como o GenAI pode ser aplicado como um copiloto no local de trabalho do DOE.

1. **Resumo:** A capacidade de resumo do GenAI pode pegar uma grande quantidade de texto e resumi-lo em um formato mais curto e digerível. Embora o modelo nem sempre atenda exatamente às solicitações de resumos de caracteres ou comprimentos de palavras específicos, ele pode criar uma correspondência próxima. A função de sumarização também pode ajudar a extrair e resumir aspectos específicos de um pedaço maior de texto — por exemplo, resumindo apenas as partes de um artigo de notícias maior que mencionam uma organização ou tópico específico.

2. **Inferência:** A funcionalidade de inferência geralmente envolve fazer previsões ou resolver problemas. Exemplos da funcionalidade de inferência do GenAI incluem pedir ao modelo para inferir o sentimento de um determinado pedaço de texto (por exemplo, sentimento positivo ou negativo) ou para fazer uma inferência sobre se há um tipo específico de informação dentro do texto (por exemplo, a marca de um item em uma revisão do produto ou se um artigo contém referências a uma entidade governamental específica). Observe que as funcionalidades de inferência carregam um conjunto específico de riscos. A análise de texto e a inferência, especificamente se essas inferências se relacionam a um indivíduo específico, devem ser usadas com considerável cautela e apenas em cenários específicos. Todos os sistemas de informação que contêm informações pessoais devem ter um concluiu a Avaliação de Impacto de Privacidade (PIA) no registro. Uma das perguntas mais complexas feitas durante uma PIA é se o sistema adicionará (criará, adquirirá ou inferirá) informações sobre a pessoa que não foram coletadas diretamente e não são oficialmente parte do registro. Os recursos de inferência também podem categorizar incorretamente ou caracterizar incorretamente as visões ou declarações feitas por indivíduos, e qualquer saída deve ser revisada por um copiloto humano.

3. **Transformação:** os modelos de texto para texto GenAI podem transformar texto de várias maneiras. A tradução é uma aplicação, pois os modelos GenAI geralmente estão familiarizados com centenas de idiomas em vários graus de proficiência. O texto pode ser traduzido para vários idiomas simultaneamente e ajustado com base na formalidade e no público-alvo. O modelo também pode transformar um pedaço de texto para refletir um novo tom ou público, como transformar uma saudação casual em um memorando comercial formal. O texto também pode ser editado para gramática e ortografia. Além disso, o texto pode ser transformado em outro formato, incluindo linguagens de codificação, como alterar um bloco de entrada de JavaScript Object Notation (JSON) para Hypertext Markup Language (HTML).

4. **Expansão:** A quarta função do GenAI de texto para texto é expandir um determinado trecho de texto ou tópico, adicionando ou criando conteúdo ou fornecendo informações adicionais sobre uma área de interesse.

Exemplos de expansão incluem usar o GenAI para escrever uma resposta a uma consulta constituinte com base no assunto e no sentimento da consulta ou para escrever um ensaio ou artigo de formato mais longo sobre um tópico fornecido pelo prompt. Observe que as funcionalidades de expansão são mais suscetíveis a riscos relacionados a direitos autorais e preocupações com propriedade intelectual (consulte a Seção 8.7) e a alucinações de IA (consulte a Seção 8.10).

6. Oportunidades para aplicar IA generativa

6.1 Principais aplicações

Ao considerar as oportunidades para casos de uso e aplicações GenAI, tenha em mente que as soluções GenAI são multimodais e podem gerar texto, imagem, áudio, vídeo e código de programação. Uma variedade de casos de uso para cada forma de mídia (modalidade) já está sendo adotada. Dentro de cada uma das modalidades, há vários casos de uso viáveis que podem ser potencialmente aplicados no DOE. A Figura 5 fornece vários dos aplicativos GenAI mais adequados para quatro modalidades.

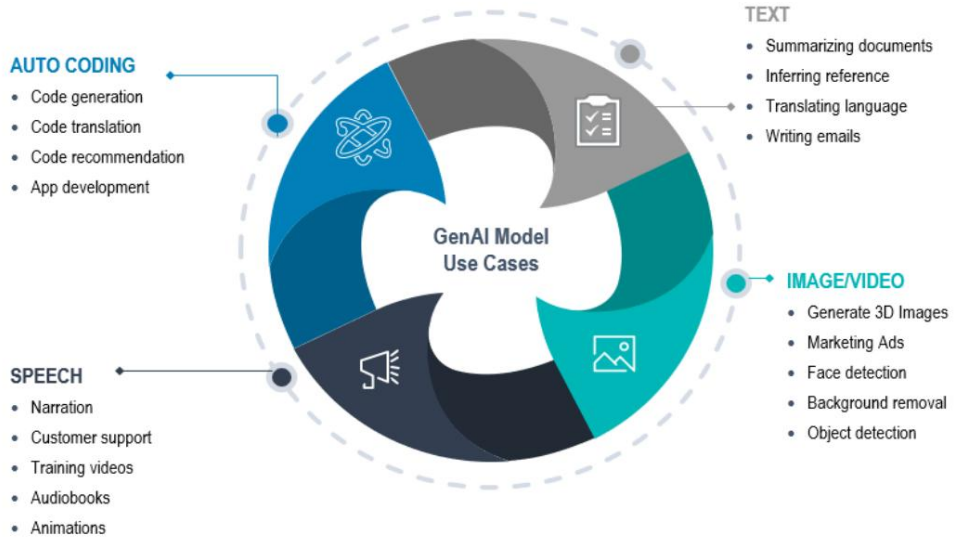


Figura 5: Principais aplicações do GenAI para texto, fala, imagem/vídeo e código

6.2 Casos de uso para DOE (Ilustrativo)

A tabela abaixo expande ideias para casos de uso categorizados por modalidade e inclui exemplos de onde o GenAI pode ser usado no DOE. Para uma visão atualizada e mais detalhada dos casos de uso de IA sendo aplicados no DOE, consulte o DOE 2023 AI Use Case Inventory (inclui casos de uso para vários recursos de IA, principalmente no espaço de análise e pesquisa de dados, e não se limita ao GenAI). Casos de uso adicionais para outros aplicativos GenAI podem se tornar disponíveis conforme o inventário amadurece. Para vários exemplos científicos, consulte o Apêndice F. De acordo com a Ordem Executiva 14110 sobre o *Desenvolvimento e Uso Seguro, Protegido e Confiável de Inteligência Artificial,* o Diretor do Escritório de Gestão e Orçamento (OMB) emitirá instruções ao DOE e outras agências federais para a coleta, relatórios e publicação de casos de uso de IA da agência anualmente, em alinhamento com a Seção 7225(a) do *Advancing American AI Act.* 11 Em relação à implementação desses e de outros casos de uso, a Seção 10.1(f)(i) da Ordem Executiva 14110 afirma que, com as proteções apropriadas em vigor, é recomendado que o acesso seja fornecido a "recursos GenAI seguros e confiáveis, pelo menos para o uso de experimentação e tarefas de rotina que não tenham impacto sobre direitos".

Exemplos de casos de uso de IA generativa

Funcionalidades de texto (por exemplo, sumarização, inferência, expansão, transformação)	
Resumo	Resumir contratos, propostas, relatórios, comentários de partes interessadas e documentos regulatórios
	Crie ou aprimore ferramentas de pesquisa interna
Inferência	Realizar análise de sentimentos a partir de uma interação, como um e-mail (por exemplo, sentimento positivo ou negativo)
Expansão	Crie primeiros rascunhos de contratos, rascunhos, apresentações comerciais, memorandos, e-mails, respostas a perguntas e solicitações de propostas (RFPs) otimizadas
	Fornecer conselhos ou informações adicionais sobre um tópico
Transformação	Traduzir documentos, contratos e comunicações para um ou mais idiomas
	Auxiliar na escrita de código de programação e documentação
	Avaliar e identificar erros no código
	Traduzir código de uma linguagem de programação para outra
	Executar autocompletar código
Funcionalidades de imagem (por exemplo, geração/criação, interpretação)	
Geração/criação (por exemplo, texto para imagem ou imagem para imagem)	Crie uma imagem com base em uma descrição de texto
	Crie um visual para um produto, campanha, página de capa, boletim informativo, logotipo, material promocional
Interpretação (por exemplo, imagem para texto)	Crie uma descrição de um visual usado em uma apresentação, por exemplo, reconheça que a imagem é uma representação de um sistema e use a legenda visual como parte da descrição visual
Funcionalidades de áudio (por exemplo, conversão de voz em texto, conversão de texto em voz, criação/geração de áudio)	
Transcrição (por exemplo, conversão de fala em texto)	Transcrever vídeos de recursos de aprendizagem para consumo como texto
	Transcrever atas de reunião
Geração/criação (por exemplo, conversão de texto em fala)	Crie uma narração de áudio para um treinamento educacional
	Gere sons personalizados ou clipes de áudio
Edição de áudio (por exemplo, conversão de voz em voz)	Edite um clipe de áudio sem precisar regravá-lo
	Traduzir a fala existente em um clipe de áudio ou vídeo para um idioma diferente usando um Voz gerada por IA ou a voz do locutor no áudio existente

Funcionalidades de vídeo (por exemplo, interpretação (vídeo para texto), criação/geração)	
Interpretação (por exemplo, vídeo para texto, fala para texto)	Revise o vídeo usado em uma proposta ou em uma reunião onde o vídeo está incluído e forneça um resumo do vídeo
	Analisar vídeos para identificar vulnerabilidades e alertar a segurança (no contexto de soluções de segurança que usam câmeras)
Criação/geração (por exemplo, texto para vídeo e/ou imagem para vídeo)	Crie vídeos para materiais de treinamento ou apresentações, potencialmente combinados com o uso de Avatares de IA
Abreviação/ condensação/ tradução (por exemplo, vídeo para vídeo)	Crie um trailer (um vídeo curto) para resumir ou abreviar um vídeo mais longo
	Use um vídeo existente para gerar o mesmo vídeo em outros idiomas
Casos de uso potenciais selecionados do GenAI	
Gerar perguntas de entrevista (por exemplo, expansão de texto)	
Caso de uso	Crie um primeiro rascunho de perguntas de entrevista para avaliação de candidatos com base em uma determinada descrição de cargo
Considerações	Avaliar o primeiro rascunho produzido pela GenAI para garantir o alinhamento com o propósito pretendido da entrevista
Crie atas de reunião (por exemplo, transcrição de áudio)	
Caso de uso	Gerar atas de reunião escritas para uma reunião do DOE a partir de uma gravação de áudio da reunião
Considerações	Divulgar aos participantes que a reunião está sendo gravada para gerenciar riscos legais e éticos
Melhore os vídeos informativos (por exemplo, personalização de vídeo)	
Caso de uso	Use o GenAI para aprimorar vídeos informativos adicionando narração de voz, gráficos, legendas ou traduções
Considerações	Semelhanças pessoais só podem ser usadas com o devido consentimento legal. No entanto, há riscos éticos e legais significativos em torno da criação e divulgação de deepfakes. Qualquer adição de apresentadores deve ser sintética (não uma “semelhança” de qualquer pessoa) a menos que tenha havido colaboração significativa com o assunto e especialistas legais. Narração de voz, traduções e legendas devem ser verificadas quanto à correção e completude.

7. Operacionalização

7.1 Operacionalização em resumo

Esta seção fornece conhecimento fundamental sobre três conceitos-chave em torno do GenAI antes de explorar as principais considerações e práticas recomendadas na Seção 8. Os três conceitos introduzidos nesta seção são funções organizacionais, dados públicos versus não públicos e modelos de serviço.

- ÿ Diferentes funções em toda a organização têm responsabilidades e considerações específicas quando se trata de chega ao GenAI.
- ÿ Como uma prática recomendada para mitigar riscos de privacidade e segurança, os usuários não devem inserir dados não públicos (sensíveis) em um sistema GenAI, a menos que os processos apropriados tenham sido realizados para garantir que os direitos e usos potenciais dos dados sejam permitidos, ou que estejam usando uma ferramenta que esteja apropriadamente configurada e aprovada para seu caso de uso. Essa prática recomendada é crítica para sistemas públicos ou comerciais onde o modelo, as entradas e as saídas não estão sob o controle direto do DOE.
- ÿ Há várias maneiras de abordar modelos de serviço. A chave é determinar se o DOE controla ou não o modelo GenAI e as saídas e se as entradas são adicionadas aos dados de treinamento do modelo. Considerações específicas se aplicam a ambos os casos.

7.2 Funções organizacionais

Todos têm um papel importante a desempenhar ao considerar e implementar uma nova solução GenAI ou usar uma ferramenta GenAI existente. Seja um usuário geral ou um especialista em sistemas de IA, cada funcionário deve estar ciente de sua função e de quaisquer considerações específicas que possam se aplicar à sua função relacionada ao desenvolvimento e uso de tecnologias GenAI. Abaixo está um conjunto introdutório de funções em toda a organização com descrições correspondentes. Observe que esta lista não é exaustiva e que, em muitos casos, essas funções usam uma linguagem específica do DOE, mas podem ter aplicações em outras organizações. Considere desenvolver uma Matriz RACI (Responsible, Accountable, Consultad, Informed) para definir claramente as funções e responsabilidades para cada solução GenAI específica. As descrições listadas abaixo são responsabilidades a serem consideradas ao redigir requisitos mais explícitos, não requisitos em si. Para obter informações adicionais sobre o Ciclo de Vida da IA referenciado nesta tabela, consulte o Apêndice G.

Muitas dessas funções ainda estão se desenvolvendo dentro do DOE. Por exemplo, as funções e responsabilidades do Chief Artificial Intelligence Officer (CAIO) e do Responsible Artificial Intelligence Officer (RAIO) variam entre as organizações. Em algumas organizações, uma pessoa pode assumir as funções de CAIO e RAIO, enquanto no DOE, essas são atualmente duas funções distintas e emergentes (no momento da publicação deste documento). No DOE, o RAIO se reporta ao Chief Intelligence Officer (CIO), enquanto o CAIO se reporta ao Secretary of Energy.

Nome da função organizacional	Descrição
Usuário geral	Independentemente de uma pessoa poder exercer uma das funções específicas listadas abaixo, quase qualquer pessoa no DOE pode ser ou em breve pode se tornar um usuário geral do GenAI. Para uso geral, é essencial entender a natureza da informação inserida no modelo, a finalidade pretendida do modelo e quaisquer restrições relacionadas aos dados de entrada ou à função de um usuário na organização. Quando necessário, relate problemas observados de precisão, imparcialidade ou viés na saída de um modelo. O usuário geral também pode incluir o humano no loop para verificar as saídas para garantir tanto a responsabilidade (ética e viés limitado) quanto a precisão, especialmente quando um sistema GenAI gera saídas para humanos consumirem, toma medidas motivadas por saídas ou tira conclusões com base em saídas.

Nome da função organizacional	Descrição
Desenvolvedor de IA	O desenvolvedor de IA é encarregado de projetar, codificar e melhorar iterativamente novos aplicativos GenAI em colaboração com outras funções, como cientistas de dados, designers de experiência do usuário, especialistas em segurança cibernética, patrocinadores de projetos e liderança. O desenvolvedor de IA cria os sistemas e soluções de IA, em oposição ao cientista de dados que desenvolve os modelos subjacentes. Os desenvolvedores de IA devem considerar as implicações únicas em torno das tecnologias GenAI e consultar as PMEs de IA conforme necessário para implementar as melhores práticas específicas de IA.
Equipe de política e governança de IA	A equipe de política e governança de IA aconselha sobre a criação de novas políticas e governança de IA com base na necessidade organizacional, técnica e legislativa para adoção e implementação em todo o Departamento de estruturas, princípios, procedimentos e práticas de IA responsáveis, éticos e confiáveis. Eles garantem que as políticas reflitam as melhores práticas em IA e abordem quaisquer preocupações de segurança, risco ou privacidade, bem como princípios de IA responsáveis e éticos. Essa função está surgindo e evoluindo muito rapidamente, e provavelmente haverá responsabilidades adicionais associadas a ela conforme a adoção do GenAI avança.
Gerente de portfólio de IA	O gerente de portfólio de IA supervisiona todos os recursos e projetos de IA atualmente em andamento para sua organização ou elemento departamental. Essa função é essencial para limitar a redundância de soluções de IA que podem ter funções semelhantes. O gerente de portfólio deve entender o cenário atual e futuro próximo da IA para identificar tendências e riscos nos recursos propostos. Essa função é responsável por todo o pipeline de todas as iniciativas de IA no nível mais alto e é responsável por todas as políticas e processos associados ao pipeline de IA. Essa função também gerencia o financiamento e o orçamento para todas as iniciativas de IA.
Especialista em IA (SME)	O especialista em IA (SME) aconselha outros sobre as melhores práticas e riscos considerações para tecnologias GenAI. Esta função precisa entender as tecnologias envolvidas para um determinado caso de uso e fornecer serviços de consultoria para outras funções, incluindo cientistas de dados, liderança e engenheiros de dados para compartilhar conhecimento com os membros apropriados da equipe. Esta função pode estar envolvida em qualquer estágio do ciclo de vida da IA. Por exemplo, o SME pode estar envolvido no estágio de planejamento inicial para garantir que um determinado problema de negócios seja adequado para uma solução GenAI, durante os estágios de desenvolvimento ou implementação para garantir eficiência e qualidade técnica ou de processo, ou fornecendo insights sobre como educar melhor os usuários sobre o uso responsável do GenAI.
Analista de negócios	O analista de negócios é encarregado de coordenar esforços entre equipes de projeto ou desenvolvimento para projetar, lançar e operar recursos GenAI. Esta função é de um usuário interno que é responsável pela tradução e coordenação de necessidades e tarefas entre os usuários de negócios e a equipe de desenvolvimento de IA/ML. Esta função traduz as necessidades de negócios em requisitos técnicos e ajuda os usuários de negócios a usar efetivamente a saída conforme o sistema foi projetado. O analista de negócios também pode identificar potenciais casos de uso comercial para tecnologias de IA ao longo de suas responsabilidades diárias.
Chefe de Inteligência Artificial Oficial (CAIO)	O Diretor de Inteligência Artificial (CAIO) é uma função definida na Ordem Executiva 14110, Seção 10.1(b)(i), que encarrega o Diretor de IA de “coordenar o uso de IA por sua agência, promover a inovação de IA em sua agência, gerenciar riscos do uso de IA por sua agência e executar as responsabilidades descritas na Seção 8(c) da Ordem Executiva 13960 (“Promovendo o Uso de Inteligência Artificial Confiável no Governo Federal”) e na Seção 4(b) da Ordem Executiva 14091”. Eles também são encarregados de supervisionar os Conselhos de Governança de IA conforme exigido por sua agência, supervisionar atividades de gerenciamento de risco para usos governamentais de IA e fazer recomendações para

Nome da função organizacional	Descrição
	agências para reduzir barreiras ao uso responsável de IA, incluindo barreiras específicas de IA de adoção para infraestrutura de tecnologia da informação, dados, força de trabalho, restrições orçamentárias e processos de segurança cibernética. Dado que esta é uma função emergente, o processo de governança e as responsabilidades para esta função ainda estão em desenvolvimento no DOE.
Oficial contratante	O oficial de contratação facilita a compra de ferramentas, plataformas, tecnologia e serviços GenAI para o DOE e usa seu conhecimento para garantir que as ferramentas GenAI mais adequadas sejam selecionadas. Essa função garante que as novas ferramentas adquiridas para o DOE sejam rigorosamente testadas, atendam aos requisitos de política e segurança, não sejam duplicadas de outros esforços contínuos na organização e estejam alinhadas com quaisquer políticas, procedimentos e estratégias existentes para a organização. Os profissionais de contratação também se preocupam com os detalhes de novos e existentes acordos contratuais e termos de serviço (ToS) com provedores terceirizados de soluções e serviços GenAI para garantir que o risco seja apropriadamente compartilhado entre o DOE e os provedores de serviço. Esta função também colabora com especialistas jurídicos/assessoria jurídica e stakeholders organizacionais.
Especialista em segurança cibernética	O especialista em segurança cibernética é responsável pela segurança, proteção e resiliência dos sistemas organizacionais (ou relacionados). Essa função deve estar envolvida desde o início de qualquer iniciativa GenAI para garantir que o design da solução tenha medidas de proteção suficientes em vigor e não interfira nas medidas de segurança existentes como resultado dos requisitos do projeto. Essa função garante que qualquer solução planejada e projetada tenha uma alta probabilidade de ser operacionalizada com sucesso. Em cada fase do ciclo de vida da IA, essa função precisa garantir que os recursos de segurança sejam mantidos adequadamente. Os profissionais de segurança cibernética também podem se concentrar em como a IA pode ser usada para reforçar a segurança cibernética em toda a organização e para prevenir ou responder a ataques adversários em sistemas não-IA.
Engenheiro de dados	O engenheiro de dados garante que os dados apropriados estejam disponíveis para os cientistas de dados e que os dados sejam tão confiáveis, justos e livres de preconceitos quanto possível. Esta função deve ter uma compreensão da estrutura de dados, ambiente, pipeline de gerenciamento e qualidade de dados em termos de sourcing, profundidade e amplitude em dados que são usados para construir sistemas GenAI. Engenheiros de dados que estão envolvidos no treinamento de modelos recém-desenvolvidos ou adquiridos também podem auxiliar na marcação de dados e outras implementações para atender às políticas de governança de dados organizacionais, e eles têm a responsabilidade de definir e implementar o plano e a arquitetura corretos para os dados. Engenheiros de dados implementam as verificações apropriadas no pipeline de gerenciamento de dados para garantir que os padrões de qualidade (normalmente definidos em uma política) sejam atendidos em cada fase do desenvolvimento.
Cientista de dados	O cientista de dados usa dados de várias fontes para auxiliar na tomada de decisões organizacionais e chegar a conclusões relacionadas ao setor. Essa função geralmente é responsável por desenvolver modelos e deve entender em detalhes o propósito pretendido e as saídas de seus modelos para garantir a funcionalidade adequada, quaisquer considerações regulatórias ou de privacidade relacionadas a um determinado modelo ou projeto e as diferenças entre treinamento/validação e produção/ dados ao vivo. Ao desenvolver ou treinar um sistema GenAI, os cientistas de dados são responsáveis por garantir a qualidade, representatividade e ausência de viés nas saídas do conjunto de dados de treinamento. O cientista de dados deve explorar os dados fornecidos pelos engenheiros de dados e aplicar as melhores metodologias e ferramentas para atingir o objetivo do projeto com os dados fornecidos. Os cientistas de dados podem assumir outras funções de

Nome da função organizacional	Descrição
	dentro desta lista, incluindo arquiteto de solução, desenvolvedor, AI SME e mais. Eles também podem assumir o papel de engenheiro rápido no caso do GenAI.
Operações de desenvolvimento Engenheiro (DevOps)	O engenheiro DevOps é responsável pelos processos que ajudam o DOE a melhorar a eficiência do desenvolvimento, teste, operacionalização e atualização de tecnologia. Essa função ajuda a facilitar esses processos com conhecimento de tecnologia emergente, habilidades de gerenciamento de projetos e comunicação de equipe. O engenheiro de DevOps colabora com outras funções técnicas para garantir a funcionalidade durante a transferência da solução do piloto para o ambiente de produção, enquanto monitora riscos e vulnerabilidades potenciais (por exemplo, desvios de dados, desvios de software ou modelo e problemas de segurança). O engenheiro de DevOps também é responsável pelo gerenciamento contínuo de modelos/sistemas, dependendo da política organizacional ou dos princípios orientadores. Essa função pode fazer parte de uma equipe maior que inclui uma variedade de conjuntos de habilidades dentro do guarda-chuva maior do DevOps, por exemplo, arquiteto de construção, gerente de lançamento, engenheiro de infraestrutura, arquiteto de automação, entre outros.
Patrocinador executivo	O patrocinador executivo é uma posição de liderança e é responsável por comunicar à organização e conscientizar sobre a importância de iniciativas estratégicas priorizadas. Essa função garante que os recursos estejam disponíveis para qualquer projeto GenAI priorizado e eleva a iniciativa em questão a um nível de prioridade mais alto. Essa função é responsável por garantir o financiamento inicial para o projeto em questão e por garantir que as partes interessadas apropriadas estejam envolvidas com a iniciativa e alinhadas com seus objetivos. O patrocinador executivo pode ocasionalmente preencher a função de diretor de projeto, que é mais prático por natureza.
Profissional de tecnologia da informação (TI)/sistemas	O profissional de tecnologia da informação é encarregado de supervisionar os sistemas de TI em toda a organização. O foco dos profissionais de TI está tanto na implementação quanto na manutenção da nova tecnologia GenAI e no fornecimento de recomendações e melhores práticas para o uso do GenAI para usuários em potencial. Essa função tem responsabilidades diferentes dependendo do estágio do ciclo de vida da IA. Por exemplo, durante a fase de desenvolvimento, o profissional de TI ajuda a construir o sandbox.
Liderança	Há muitas responsabilidades potenciais diferentes que podem pertencer à função de liderança. A liderança define a direção estratégica, prioridades, metas e objetivos de missão para a organização em colaboração com autoridades federais. A liderança está preocupada em entender o GenAI no nível executivo, entender o amplo cenário regulatório no que se refere ao GenAI e construir conscientização e compreensão do GenAI e seus casos de uso em toda a organização. A liderança no nível do Departamento e no nível do Elemento Departamental (DE)/local garante que o treinamento sobre o uso, limitações e riscos do GenAI esteja disponível e seja incentivado para todos os usuários potenciais do GenAI no DOE. A liderança também facilita a criação de um mecanismo central e colaborativo para compartilhar conhecimento, colaborar em iniciativas, relatar observações de preconceito, falta de confiabilidade, problemas de segurança e outras preocupações nas plataformas GenAI como um meio para o aprendizado organizacional. Este grupo também pode incluir líderes que não interagem diretamente com tecnologias de IA.
Especialista jurídico em IA e tecnologia emergente	Profissionais jurídicos especializados em tecnologia emergente e IA navegam no intrincado reino da IA na lei, focando nas implicações e requisitos exclusivos da tecnologia emergente e da IA. Eles são encarregados de compreender as nuances dos modelos de IA, utilizando-os efetivamente para tarefas legais e alinhando essas aplicações com padrões éticos e estruturas legais estabelecidas. Eles garantem que o uso de IA em operações jurídicas não apenas otimiza a eficiência e a precisão, mas também mantém protocolos de privacidade robustos,

Nome da função organizacional	Descrição
	aborda riscos potenciais e adere às normas e regulamentações legais. Especialistas jurídicos são responsáveis por se manterem atualizados sobre a legislação existente e pendente para proteger proativamente a agência e se preparar para o que está por vir. Eles também servem como consultores para toda a equipe da GenAI.
Gestão e operação Pessoal (M&O)	A equipe contratada da M&O supervisiona os Laboratórios Nacionais e precisa entender o riscos legais, técnicos e de aquisição assumidos como um terceiro organizacional ao monitorar o desempenho das soluções GenAI em operação. Esta função se concentra na manutenção, monitoramento e uso geral do sistema GenAI uma vez implantado. Nos estágios de design e desenvolvimento, esta função garante que os requisitos sejam realistas e considerados na construção da solução. Observe que a equipe de M&O inclui uma ampla gama de pessoas, incluindo pesquisadores, executivos, especialistas em segurança cibernética e muito mais. Muitas das funções nesta lista podem se aplicar a funcionários de M&O selecionados.
Gerente de produto	O gerente de produto é responsável por entender e priorizar oportunidades de mercado relevantes para a organização para casos de uso de IA, em colaboração com os proprietários de casos de uso e a liderança. Essa função é responsável por gerenciar as atividades e a equipe para a iniciativa, produto ou serviço GenAI específico em questão. Pode haver uma variedade de gerentes de produto conforme o portfólio de IA se expande.
Gerente de programa	O gerente de programa é uma função única associada às instituições de pesquisa. Esta função escreve anúncios de oportunidades de financiamento, analisa propostas, faz recomendações de financiamento e gerencia prêmios, entre outras responsabilidades. Para o Office of Science, esses prêmios são principalmente para universidades e Laboratórios Nacionais e focam em pesquisa fundamental. Embora propostas e prêmios possam incluir o desenvolvimento e o uso de IA para ciência e engenharia, também há uma oportunidade para a GenAI auxiliar o gerente do programa no desempenho de suas funções.
Diretor do projeto	O diretor do projeto é encarregado de comprometer tempo e recursos para supervisionar e revisar o desenvolvimento de novos recursos GenAI em um ambiente baseado em projeto, o que inclui tomar decisões de alto nível sobre a direção e os objetivos do projeto, como o projeto é gerenciado e estruturado e como as tecnologias GenAI podem contribuir para outras iniciativas em andamento. Os diretores de projeto podem ter sobreposição com a função de liderança. Os diretores de projeto fornecem requisitos diretos dependendo do caso de uso e garantem que todas as funções na equipe GenAI sejam convidadas a colaborar e que as iniciativas sejam concluídas em alinhamento com a estratégia geral.
Cientista pesquisador	O cientista pesquisador investiga pesquisas de código aberto e publicadas e/ou conduz estudos e experimentos. Cientistas pesquisadores estão especialmente preocupados com a precisão técnica da saída ao utilizar uma ferramenta GenAI em seu trabalho, bem como entender as considerações de direitos autorais e publicação para qualquer pesquisa feita usando uma ferramenta GenAI. O uso responsável e ético das ferramentas GenAI é uma preocupação fundamental para cientistas pesquisadores. Observe que os papéis de muitos cientistas pesquisadores do DOE diferem em seu avanço da ciência no uso de IA. Consulte o Apêndice F para referências a recursos adicionais sobre P&D do DOE e avanço da ciência. No momento, pesquisadores que analisam bolsas como parte do processo de revisão de mérito da NSF não podem usar IA como um auxílio em nenhuma capacidade.12
Oficial Responsável pela IA (RAIO)	O RAIO é responsável por gerenciar um programa de gerenciamento de risco de IA, colaborando com os funcionários apropriados para estabelecer ou atualizar processos para avaliar o desempenho dos sistemas de IA, supervisionando a conformidade do DOE com os requisitos para gerenciar riscos de IA e conduzindo avaliações de risco quando

Nome da função organizacional	Descrição
	necessário. O RAIO também é responsável por coordenar a implementação dos nove Princípios de IA Confiáveis estabelecidos na Seção 3 da EO 13960. Dado que esta é uma função emergente, o processo de governança e as responsabilidades para esta função, bem como sua implementação, ainda estão em desenvolvimento no DOE.
Arquiteto de soluções	O arquiteto de soluções supervisiona a integração das tecnologias GenAI na infraestrutura geral de TI organizacional. Isso pode exigir a implementação de medidas adicionais de governança ou segurança de dados, bem como uma conscientização sobre os tipos de fluxos de dados e acesso que são ou não permitidos com as tecnologias GenAI. Essa função ajuda a projetar como o sistema ficará e funcionará com base nos requisitos fornecidos e na finalidade pretendida. Essa função pensa em como o modelo GenAI será integrado e operacionalizado com os sistemas upstream e downstream e garante que tudo corra bem quando o POC bem-sucedido for movido para a produção.
Caso de uso proprietário da empresa	O proprietário do negócio do caso de uso é responsável por estabelecer um caso de negócio para uma solução de IA nova ou expandida, comunicar as necessidades do negócio à equipe de desenvolvimento e/ou aquisição, colaborar com outras funções para selecionar uma solução para o caso de uso e auxiliar na implementação e manutenção da solução GenAI conforme necessário. Essa função está envolvida diariamente com o caso de uso fornecido e pode ser responsável por fornecer os dados para o desenvolvimento da solução.
Designer de experiência do usuário (UX)	O designer de experiência do usuário (UX) é responsável por criar a interface do usuário centrada no ser humano de uma solução tecnológica ou produto, incluindo o design dos componentes e recursos que controlam como um usuário interage com o produto. Os designers de UX se concentram em como adaptar a interface do usuário para atender às necessidades dos usuários finais-alvo da solução, para melhorar a qualidade da experiência do cliente e para tornar a interface do usuário o mais simples e eficaz possível para o usuário final-alvo do produto. É essencial para essa função incorporar uma abordagem de design centrada no ser humano na interface do usuário no produto final.

7.3 Dados públicos vs. dados protegidos

Com todas as formas de IA, e especialmente no contexto de GenAI, há considerações críticas em torno do uso de dados e informações. Dados (ou seja, informações registradas, independentemente da forma ou da mídia em que podem ser registradas, incluindo dados técnicos e software de computador) podem ser amplamente categorizados em duas classes: públicos e protegidos.

- ÿ Dados protegidos incluem informações que são protegidas contra distribuição pública e/ou certos usos, e inclui dados sensíveis, dados privados, dados proprietários, dados confidenciais e trabalhos protegidos por direitos autorais. Dados confidenciais são explorados mais detalhadamente na Seção 8.6 e não devem ser confundidos com informações classificadas que também foram marcadas como confidenciais. Para definições de vários tipos de dados sensíveis, consulte o Apêndice J: Exemplos de Dados Protegidos. Alguns tipos de dados protegidos são protegidos e/ou não públicos apenas por um período de tempo definido, enquanto outros tipos de dados protegidos podem permanecer protegidos e não públicos indefinidamente.

- ÿ Dados públicos são informações que podem ser livremente utilizadas e distribuídas por qualquer pessoa, sem restrições legais quanto ao acesso ou uso.

A principal lição é que nenhum tipo de ***dado protegido ou informação não pública deve ser compartilhado ou inserido em nenhum sistema GenAI público ou comercial (não controlado pelo DOE).*** O DOE controla o sistema GenAI se for um sistema de IA fechado ou proprietário. Um sistema de IA fechado é desenvolvido e controlado

por uma única organização que tenha controle total e propriedade sobre o sistema. Exceções podem ser aplicadas se o sistema GenAI for protegido por um acordo de confidencialidade entre o DOE e o fornecedor e o DOE tiver direitos suficientes para usar os dados protegidos ou não públicos.13 Esta recomendação é uma prática recomendada crítica para o uso do GenAI pelo DOE. Mesmo para sistemas GenAI controlados pelo DOE, os usuários devem garantir que tenham os direitos apropriados para usar quaisquer dados não públicos ou protegidos que estejam sendo utilizados.

A principal lição para dados públicos é que a possibilidade de a informação publicamente disponível inserida no sistema GenAI ser incorporada ao modelo de IA pode não representar uma ameaça, já que os dados já são públicos. No entanto, é importante lembrar do risco de plágio e violação de direitos autorais e verificar dados publicamente disponíveis, mesmo que eles possam ser usados livremente.

Existem muitas leis, mandatos, políticas e treinamentos internos do DOE que fornecem proteções valiosas sobre o uso geral de dados, especialmente políticas sobre proteção e compartilhamento de informações.
Isso deve receber atenção especial ao usar o GenAI.

Recursos adicionais sobre dados incluem:

- ÿ Recursos do DOE: CUI Slicksheet, Informações não classificadas controladas
- ÿ "Programa de Informação de Infraestrutura Crítica Protegida (PCII)", Segurança Cibernética e Agência de Segurança de Infraestrutura (CISA)
- ÿ Regulamentação Federal de Aquisições (FAR), Acquisition.gov

O uso ou divulgação indevida e não autorizada de dados protegidos ou informações não públicas pode levar à responsabilidade legal tanto para o DOE quanto para o indivíduo responsável pelo uso não autorizado da divulgação. Em alguns casos, a responsabilidade para indivíduos pode incluir potenciais penalidades civis e criminais.

7.4 Modelos de serviço

O uso de soluções GenAI pode apresentar uma variedade de riscos, dependendo de quem tem o controle da solução em uso. A principal distinção se resume à seguinte questão: O DOE **controla** ou **não** a solução? (Consulte a Seção 7.3 diretamente acima para uma breve discussão sobre a distinção entre um sistema controlado pelo DOE e um sistema não controlado pelo DOE.) Sistemas prontos para uso e que não foram adquiridos e personalizados pelo DOE ou acessíveis publicamente (por exemplo, ChatGPT, Bard) apresentarão um conjunto maior de riscos em comparação a um aplicativo criado internamente e protegido dentro dos limites do DOE. É importante entender o tipo de aplicativo em uso e cumprir as políticas e diretrizes existentes do DOE quando se trata do uso de tecnologia. Para esclarecer a instância de um aplicativo que deve ser usado ou para quaisquer outras questões relacionadas a sistemas, entre em contato com o Responsible AI Official (RAIO), o Chief AI Officer (CAIO), a equipe de Supply Chain Risk Management (SCRM) do OCIO ou a equipe local de Tecnologia da Informação (TI). A compreensão e a documentação sólidas das informações detalhadas abaixo são essenciais para implementar as melhores práticas listadas na Seção 8.

A seguir estão algumas questões a serem consideradas antes de usar um aplicativo GenAI:

- ÿ Quais são os benefícios específicos que o uso desta solução proporcionaria?
- ÿ Quem construiu o aplicativo e/ou modelo?
- ÿ Esta solução é desenvolvida e/ou controlada pelo DOE? Se não, há uma ferramenta que seja?
- ÿ Onde o aplicativo está hospedado?
- ÿ Como o modelo foi treinado?
- ÿ Como os dados foram selecionados e coletados?
- ÿ Quais dados foram usados para treinar o modelo e em que data?

- O uso é permitido sob os direitos de dados disponíveis, dada a fonte dos dados ou para o tipo de Informação?
- Qual modelo, plataforma e metodologia foram utilizados?
- Este é um aplicativo público, está em uma nuvem privada segura ou está operando dentro de uma nuvem segura do DOE? rede?
- O aplicativo é um produto pronto para uso?
- Quando as informações são fornecidas ao aplicativo, qual o risco das informações se tornarem públicas?
- Como o modelo foi validado?
- Os termos de serviço do sistema GenAI são “federalmente compatíveis” com as diretrizes estabelecido pela Administração de Serviços Gerais (GSA)? Além disso, os termos de serviço para tratamento de dados atendem aos requisitos apropriados de privacidade ou confidencialidade?
- Quais são os limites do risco compartilhado ou transferido e da responsabilidade entre o DOE e todos os envolvidos? festas?

Para sistemas criados internamente para o DOE que processam rotineiramente informações comerciais do DOE, soluções locais ou híbridas, as seguintes considerações adicionais são necessárias antes de operacionalizar o GenAI:

- O aplicativo foi aprovado pelos padrões e procedimentos de segurança cibernética do DOE (por exemplo, A Autorização para Operar (ATO) foi emitida, a Avaliação de Impacto à Privacidade (PIA) foi implementada, etc.)?
- O aplicativo é aprovado pelo FedRAMP?
- O modelo é totalmente independente, sem recuperação de terceiros?
- Existe um contrato ou acordo de serviço em vigor?

8. Principais considerações e melhores práticas

8.1 Principais considerações e melhores práticas em resumo

Esta seção fornece uma visão geral de sete considerações onde certos riscos são conhecidos por surgir com tecnologias GenAI. Cada uma das sete subseções fornece uma breve descrição da consideração, exemplos públicos e ilustrativos de onde e como os riscos podem surgir, riscos específicos e melhores práticas para mitigar esses riscos. Os sete tópicos abordados nesta seção são segurança e resiliência, privacidade, confidencialidade, propriedade intelectual, segurança, justiça e preconceito, e alucinações e interpretações de IA. No final desta discussão na Seção 8.11: Lista de verificação de melhores práticas, um resumo das melhores práticas é fornecido. Observe que podem surgir desafios ao implementar as melhores práticas e tenha em mente que as melhores práticas para GenAI continuarão a surgir e evoluir. Consulte o Manual de gerenciamento de risco de IA do DOE (AIRMP) para mais ideias sobre riscos relacionados à IA e estratégias de mitigação de riscos.

8.2 Introdução

Há sempre considerações específicas, riscos únicos e melhores práticas que devem receber atenção ao embarcar em uma jornada para inovar com tecnologia. A IA tem mais considerações únicas do que tecnologias não IA devido à natureza complexa dos modelos e sua dependência de conjuntos de dados. A GenAI é ainda mais complexa e, portanto, vem com considerações, riscos e estratégias de mitigação de riscos ainda mais matizados. Toda a organização precisa entender a natureza complexa dos riscos da GenAI e as melhores práticas para maximizar os benefícios da GenAI enquanto minimiza seus riscos.

Cada função listada na Seção 7.2: Funções Organizacionais, incluindo usuários gerais, tem considerações específicas do GenAI e melhores práticas associadas a ela. É essencial criar conscientização em toda a organização sobre essas funções, responsabilidades e melhores práticas.

Esta seção fornece detalhes sobre sete considerações-chave e melhores práticas relativas ao GenAI. Seção 8.3: Gerenciamento de Riscos de IA apresenta as sete características de sistemas de IA confiáveis descritas no National Institute of Standards and Technology (NIST) Artificial Intelligence Risk Management Framework (NIST AI RMF 1.0). Essas sete características são usadas como uma estrutura para discutir sete considerações-chave em torno do GenAI, cada uma das quais tem sua própria subseção (Seções 8.4 – 8.10).

8.3 Gerenciamento de Riscos de IA

Ao desenvolver e implementar novas tecnologias GenAI ou ao incorporar funcionalidades GenAI em sistemas existentes, é essencial entender as considerações de gerenciamento de riscos existentes e os riscos exclusivos associados ao GenAI. O GenAI introduz uma camada adicional de considerações de risco, incluindo alucinações, interpretações errôneas, envenenamento por treinamento, injeção imediata, deepfakes e violação de propriedade intelectual. Tenha em mente que riscos *não* específicos do GenAI podem se tornar mais pronunciados quando o GenAI é integrado ao ecossistema de tecnologia. É melhor projetar sistemas GenAI para serem seguros, responsáveis e confiáveis no início de qualquer iniciativa GenAI, e o gerenciamento de riscos de IA eficaz é um componente crítico para atingir essas metas. Quando empregado adequadamente, o gerenciamento de riscos de IA também permite que usuários e desenvolvedores entendam as limitações e ambiguidades da IA e habilitem a seleção de casos de uso de IA apropriados, responsáveis e viáveis.

O gerenciamento de risco de IA difere de várias maneiras das práticas de gerenciamento de risco de tecnologia não IA. A governança do programa de gerenciamento de risco normalmente inclui um conjunto de métricas para medir o desempenho e o progresso com base em dados públicos e históricos. No entanto, os casos de uso de IA geralmente não têm métricas confiáveis para usar em comparação, e as métricas podem não capturar totalmente os fatores ou impactos relevantes. Também há uma falta de consenso sobre como definir métricas claras para confiabilidade ou confiabilidade em sistemas de IA.

Outra diferença é que os sistemas de IA projetados para aumentar as ações humanas (que têm critérios de gerenciamento de risco existentes) agem de forma diferente do processo de pensamento humano, o que pode tornar os riscos específicos da IA mais complexos.

requisitos de gerenciamento difíceis de operacionalizar. Finalmente, a priorização de recursos de risco de IA pode ser decidida de forma diferente do que com estratégias de gerenciamento de risco não relacionadas a IA. Métricas de priorização para sistemas de IA podem incluir aquelas que interagem com humanos, que têm efeitos posteriores na segurança ou que têm conjuntos de treinamento que incluem informações pessoalmente identificáveis (PII).15

O NIST Artificial Intelligence Risk Management Framework (NIST AI RMF 1.0) é um excelente recurso para se familiarizar com o gerenciamento de riscos relacionados à IA e práticas de IA responsáveis. O NIST AI RMF é bastante citado na última Ordem Executiva 14110, que exige que o Secretário de Energia colabore com o Secretário de Comércio, Secretário de Segurança Interna e outros para desenvolver diretrizes e melhores práticas para desenvolver e implantar sistemas de IA seguros, protegidos e confiáveis, inclusive desenvolvendo um recurso complementar ao NIST AI RMF para GenAI. As práticas de gerenciamento de riscos descritas no NIST AI RMF são consideradas o padrão atual pelo governo federal.

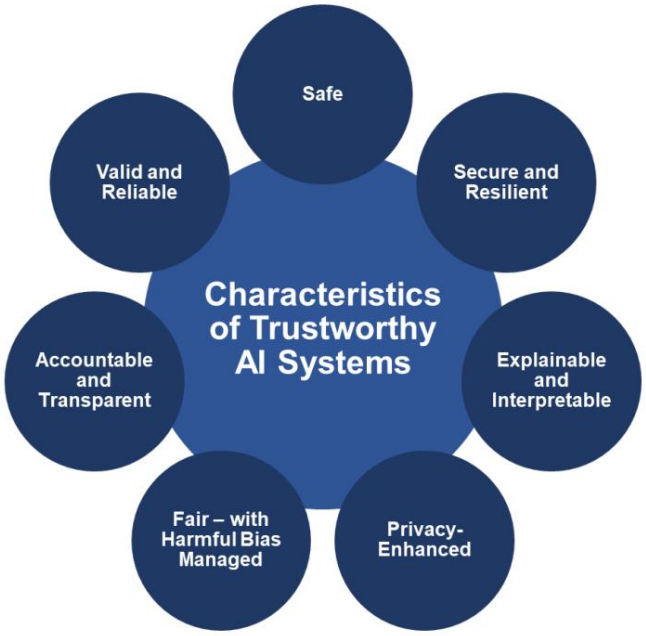


Figura 6: Sete características de sistemas de IA confiáveis delineado no NIST AI RMF 1.0

O NIST AI RMF descreve três conceitos principais a serem enfatizados no desenvolvimento de IA responsável: centralidade humana, responsabilidade social e sustentabilidade.16 Com esses conceitos principais de IA responsável em mente, o gerenciamento de riscos de IA pode permitir uso, práticas e processos responsáveis, incentivando funcionários em todo o ecossistema do DOE a praticar o pensamento crítico sobre riscos potenciais e impactos inesperados da IA.

Um tema crítico abrangente no design, desenvolvimento e implantação de IA de uma forma que maximize seus benefícios enquanto gerencia adequadamente seus riscos é a confiabilidade. IA confiável é um conceito refletido em inúmeras publicações federais relevantes, incluindo a Ordem Executiva 13960 sobre *a Promoção do Uso de Inteligência Artificial Confiável no Governo Federal* e a Ordem Executiva 14110 sobre o *Desenvolvimento e Uso Seguro, Protegido e Confiável de Inteligência Artificial.*

O NIST AI RMF lista sete características de sistemas de IA confiáveis para orientar o gerenciamento de risco de IA e o desenvolvimento responsável de IA. Essas sete características de IA confiáveis definidas pelo NIST são: Segura, Protegida e Resiliente, Explicável e Interpretável, Privacidade Aprimorada, Justa – com Viés Prejudicial Gerenciado, Válida e Confiável, e Responsável e Transparente (consulte a Figura 6). Para obter informações adicionais sobre o NIST AI RMF, consulte o Apêndice H ou a publicação online completa. As sete características de uma IA confiável, conforme descritas pelo NIST AI RMF, são usadas como uma estrutura para a discussão das principais considerações que se seguem nas Seções 8.4 – 8.10.

Nas subseções a seguir, as principais considerações e melhores práticas para GenAI são apresentadas em sete áreas de alto risco: Segurança e Resiliência, Segurança, Privacidade, Confidencialidade, Propriedade Intelectual, Justiça e Viés, e Alucinações e Interpretações Erradas e são mapeadas em relação às sete características de IA confiáveis do NIST AI RMF listadas acima. Todas as sete considerações do DOE têm implicações e aspectos confiáveis. Quando as melhores práticas correspondentes às sete principais considerações do DOE são cuidadosamente aplicadas, os sistemas GenAI serão confiáveis por design. As sete principais considerações do DOE são mapeadas em relação às sete características de IA confiáveis do NIST AI RMF da seguinte forma:

Principais considerações do DOE	Característica(s) de IA confiável(eis) NIST AI RMF 1.0
Segurança e Resiliência	Seguro e resiliente
Segurança	Seguro
Privacidade	Privacidade aprimorada
Confidencialidade	Seguro e resiliente; Seguro
Propriedade intelectual	Seguro e resiliente; responsável e transparente
Justiça e preconceito	Justo – com viés prejudicial gerenciado
Alucinações e interpretações errôneas	Responsável e transparente; Válido e confiável; Explicável e Interpretável

As seções subsequentes se aprofundam nas sete principais considerações do DOE para o GenAI mencionadas anteriormente. A Seção 8.11 fornece uma lista de verificação para resumir as melhores práticas de maior prioridade para todos os sete tópicos, bem como melhores práticas mais gerais não listadas nas sete subseções.

8.4 Segurança e Resiliência

Embora as soluções GenAI tenham surgido como ferramentas inovadoras para impulsionar a ciência, a operação e a transformação empresarial, elas também apresentam riscos de segurança que devem ser cuidadosamente abordados e mitigados.

Definição

Sistemas de IA que conseguem manter confidencialidade, integridade e disponibilidade por meio de mecanismos de proteção que impedem acesso e uso não autorizados, incluindo modificação secreta de dados de treinamento ou modelos fundamentais, podem ser considerados **seguros.**

Os sistemas de IA são considerados **resilientes** se puderem suportar eventos adversos inesperados ou mudanças inesperadas em seu ambiente ou uso — ou se puderem manter suas funções e estrutura diante de mudanças internas e externas e se degradarem com segurança e elegância quando necessário.17

Exemplos

Exemplo ilustrativo

Aplicações apoiadas por serviços GenAI que são então instalados em dispositivos que podem participar automaticamente de reuniões ou acessar outros dados e serviços. Esses aplicativos tornam as tentativas de phishing/smishing mais realistas e convincentes com imagens e voz deepfake. Tudo isso deve ser coberto nas práticas e regulamentações de segurança cibernética existentes, mas são emergentes e têm um cenário de ameaças maior.

Exemplo público

Na primavera de 2023, uma vulnerabilidade no código-fonte do ChatGPT expôs informações confidenciais dos usuários e permitiu que jogadores adversários visualizassem o histórico de bate-papo dos usuários. Alguns dos dados que foram expostos incluíam nomes, endereços de e-mail, tipos de cartão de crédito, endereços de pagamento e históricos de bate-papo. As possíveis consequências deste incidente incluem a exposição de dados privados (pertencentes a indivíduos e empresas), danos à reputação e repercussões legais.

Leia mais aqui: Primeira violação de dados da Generative AI: OpenAI toma medidas, bug corrigido | Mercados e Mercados

Considerações importantes

ÿ Informações de identificação pessoal (PII), bem como informações sensíveis, confidenciais, proprietárias ou protegidas de outra forma armazenadas por um sistema GenAI que foram inseridas como parte de um prompt ou coletadas como parte do processo de treinamento do modelo podem ser acessadas por um invasor ou outros jogadores adversários.

ÿ Jogadores adversários podem usar “injeção rápida”, um método usado por hackers que engana o sistema para contornar proteções éticas ou de segurança específicas que foram corrigidas em modelos fundamentais, para manipular os sistemas GenAI para gerar informações não autorizadas.

ÿ Ferramentas GenAI como o ChatGPT podem ser enganadas para gerar código de programação de malware ou ransomware.

ÿ Jogadores adversários podem envenenar dados para criar vulnerabilidades no sistema.

ÿ Deepfakes, ou imagens ou vídeos forçados digitalmente, podem ser criados usando o GenAI.

Melhores práticas

ÿ Monitorar e testar o sistema GenAI quanto a vulnerabilidades, ameaças, falhas, etc. e trabalhar para desenvolver métodos de teste e proteção de sistemas de IA de forma mais eficiente.

ÿ Desenvolver e implementar recursos de detecção que possam identificar ameaças, falhas e ataques no sistema e notificar o pessoal.18

- ÿ Treinar os usuários para entender os riscos de segurança associados à IA, incluindo o potencial de uso malicioso ou ataques adversários, bem como riscos à validação de entrada e saída e à integridade dos dados.
- ÿ Desenvolver e atualizar regularmente sistemas robustos e seguros para que os sites se defendam contra ameaças e planejar a resiliência do sistema para garantir que os sistemas de IA possam se recuperar de possíveis ataques ou falhas.
- ÿ Para sistemas GenAI que o DOE desenvolveu ou para os quais o DOE compilou um conjunto de treinamento especializado, estabeleça um programa para conduzir testes adversários por meio de “red teaming”, que envolve buscar ativamente exemplos de onde o sistema GenAI falha, retreinar o modelo nesses exemplos e continuar esse processo iterativo até que a equipe feche o ciclo de identificação de falhas.19
- ÿ Atualize regularmente as disposições do plano de gerenciamento de riscos do sistema para refletir os riscos mais recentes.
- ÿ Isole os sistemas GenAI tanto quanto for prático e evite permitir que os sistemas GenAI controlem diretamente outros sistemas (especialmente sistemas físicos do mundo real).

Recursos adicionais

- ÿ Estrutura de gerenciamento de risco de IA do NIST
- ÿ Manual de gerenciamento de risco de inteligência artificial do DOE

8.5 Privacidade

Proteger a privacidade é fundamental para preservar a confiança do público no governo. O governo federal se esforça para manter os mais altos padrões na coleta, manutenção, uso e disseminação de informações pessoais das pessoas.

Definição

A consideração **da privacidade** envolve proteger a segurança das informações pessoais para garantir sua precisão, relevância, atualidade e integridade, evitando divulgação não autorizada e garantindo que nenhum sistema de registros relativos a indivíduos, não importa quão insignificante ou especializado, seja mantido sem aviso público.20

Exemplos

Exemplo ilustrativo

Durante uma pandemia global com preocupações sobre contato e exposição humana, uma organização busca entender onde os recursos estão fisicamente presentes para fins de limpeza e desinfecção.
Um modelo de IA é usado para consumir e treinar dados de entrada. Conforme os dados são alimentados no modelo, a saída pode fornecer informações de uma forma que expõe involuntariamente a geolocalização dos indivíduos. Essa situação agora exige notificações legais adicionais aos indivíduos em relação à privacidade.

Exemplo público

Em 2021, autoridades canadenses de privacidade descobriram que a empresa americana Clearview AI estava coletando fotos de cidadãos canadenses, incluindo crianças, sem seu conhecimento ou consentimento para uso em um software de reconhecimento facial que era usado por agências de segurança pública para identificar pessoas de interesse ou vítimas. A IA utilizou bilhões de fotos encontradas na internet e em contas de mídia social para tentar identificar a pessoa.

Mesmo após o uso da tecnologia ter sido interrompido, a Clearview continuou a usar fotos de cidadãos canadenses. Especialistas alertaram que não apenas o armazenamento de dados era contra as leis de privacidade, mas que a ferramenta em si poderia ser facilmente mal utilizada. Essa questão também destaca o conceito de consentimento na coleta de dados de treinamento.

Leia mais aqui: A empresa de tecnologia dos EUA Clearview AI violou a lei de privacidade canadense: relatório | CBC News

Considerações importantes

ÿ As plataformas de IA e os provedores de serviços podem compartilhar informações do usuário com terceiros, incluindo fornecedores, provedores de serviços, afiliados ou outros usuários, sem informar o usuário.

ÿ As informações inseridas em um sistema GenAI podem se tornar parte de seu conjunto de dados de treinamento. Assim, qualquer Dados proprietários, sensíveis, pessoalmente identificáveis, confidenciais ou protegidos de outra forma inseridos como parte de um prompt podem ser usados em saídas para outros usuários do sistema.

Melhores práticas

ÿ Integrar considerações programáticas de privacidade em uma ampla gama de funções, incluindo, mas não se limitando a, segurança da informação, gerenciamento de registros, planejamento estratégico, orçamento e aquisição, contratados e terceiros, força de trabalho, treinamento, resposta a incidentes e gerenciamento de riscos.

ÿ Continue a aderir às políticas e procedimentos de privacidade existentes. Revise iterativamente as novas políticas de privacidade e procedimentos recomendados para garantir que sejam contabilizados nos casos de uso do GenAI. ÿ Exercite práticas de minimização de dados tomando medidas para anonimizar os dados e limitar a coleta, o armazenamento e a reutilização de informações pessoais.

ÿ Não inclua dados protegidos ou informações não públicas (como parte da entrada em qualquer sistema GenAI comercial ou aberto).

ÿ Descreva de forma clara e precisa, e compartilhe de forma acessível, como o departamento utiliza, gerencia e coleta informações. Documente claramente quem cria, contribui e tem acesso a essas informações, e comunique isso a todas as pessoas que confiam ao governo seus dados e informações.21

ÿ Implementar a administração completa do ciclo de vida dos dados, que é a prática de proteger e proteger dados, metadados e informações ao longo de seu ciclo de vida. Isso inclui coleta, armazenamento, uso, controle, processamento, publicação, transferência, retenção e disposição.22

ÿ . Desenvolver treinamentos de privacidade GenAI personalizados para funcionários com acesso a dados protegidos (por exemplo, dados sensíveis, dados privados, dados confidenciais, dados de direitos limitados, dados proprietários, etc.) ou qualquer outra informação não pública.23

ÿ Realizar uma avaliação de impacto na privacidade (uma análise de como as informações são tratadas para garantir o manuseio está em conformidade com os requisitos legais, regulamentares e de política aplicáveis em relação à privacidade, para determinar os riscos e efeitos da criação, coleta, uso, processamento, armazenamento, manutenção, disseminação, divulgação e descarte de informações em formato identificável em um sistema de informações eletrônico e para examinar e avaliar proteções e processos alternativos para manuseio de informações para mitigar potenciais preocupações com privacidade), conforme necessário ao desenvolver e implementar uma nova tecnologia.24

ÿ Garantias de privacidade diferencial, ou proteções que permitem que informações sobre um grupo sejam compartilhadas, ao mesmo tempo em que limitam comprovadamente o acesso, uso ou divulgação indevidos de informações pessoais sobre entidades específicas, devem ser compreendidas quando em vigor, incluindo como essas garantias afetam os dados compartilhados, usados para treinar ou criados por tecnologias GenAI.25

ÿ O GenAI também pode facilitar ou interagir com tecnologias de aprimoramento de privacidade (PETs), ou “qualquer solução de software ou hardware, processo técnico, técnica ou outros meios tecnológicos de mitigar riscos de privacidade decorrentes do processamento de dados, incluindo o aprimoramento da previsibilidade, gerenciabilidade, dissociabilidade, armazenamento, segurança e confidencialidade. Esses meios tecnológicos podem incluir computação multipartidária segura, criptografia homomórfica, provas de conhecimento zero, aprendizado federado, enclaves seguros, privacidade diferencial e ferramentas de geração de dados sintéticos. Isso também é às vezes chamado de “tecnologia de preservação de privacidade”. 26 Entenda como o GenAI pode ser usado para aumentar a privacidade em recursos recém-desenvolvidos ou existentes, ou como a introdução do GenAI em um sistema pode afetar os PETs existentes.

ÿ De acordo com o NIST AI RMF, “tecnologias de aprimoramento de privacidade… bem como métodos de minimização de dados, como desidentificação e agregação para certas saídas de modelo, podem dar suporte ao design para sistemas de IA com privacidade aprimorada. Sob certas condições, como escassez de dados, técnicas de aprimoramento de privacidade podem resultar em perda de precisão, afetando decisões sobre justiça e outros valores em certos domínios.”27

Recursos adicionais

ÿ DOE O 206.1 Programa de Privacidade do Departamento de Energia, 16 de janeiro de 2009

ÿ Escritório de Ciência e Tecnologia (OSTP), Projeto para uma Declaração de Direitos da IA, outubro de 2022

8.6 Confidencialidade

Conforme observado na Seção 7.3, a palavra “confidencial” *quando usada neste Guia* fica fora da definição do NSI.
As informações confidenciais inseridas em uma ferramenta GenAI podem ser armazenadas ou processadas pela ferramenta ou seus provedores, revelando informações confidenciais a pessoal não autorizado. Em documentos do governo dos EUA, a palavra “Confidencial” tem uma definição específica de Informações de Segurança Nacional (NSI) que se relaciona ao nível de gravidade do dano se um documento marcado como “Confidencial” for compartilhado de forma inadequada. Neste Guia, o uso do a palavra “confidencial” está fora do contexto do NSI. Em vez disso, ela está associada ao ambiente de negócios não governamentais.

Definição

Confidencialidade, conforme discutido neste Guia, é definida como “preservar restrições autorizadas de acesso e divulgação, incluindo meios para proteger a privacidade pessoal e informações proprietárias”.28

O DOE Operations Security Handbook afirma que há duas características primárias de uma informação que determinam se essa informação é segura para divulgação pública ou se deve ser considerada **sensível.** Essas duas características primárias para “determinar a adequação para divulgação de informações” são sensibilidade e risco.

- Sensibilidade: “Se a informação for divulgada ao público, ela não deve revelar ou identificar informações sensíveis informações, atividades ou programas.” Informações sensíveis também podem ser definidas como informações que podem ser usadas por adversários em detrimento da organização, seus funcionários, o público ou a nação. A sensibilidade mede o nível de dano que pode resultar da divulgação.
- Risco: “Informações que podem ser usadas por adversários em detrimento de funcionários, do público, do departamento ou da nação não devem ser aprovadas para divulgação. Essa determinação deve ser baseada em princípios sólidos de gerenciamento de risco focados na prevenção de potenciais consequências adversas.” Em termos da definição de sensibilidade apresentada acima, risco é a probabilidade de tal dano.

Juntas, essas duas características sugerem que o termo “informação sensível” é um termo categórico que inclui outros tipos específicos de informações sensíveis.29

Consulte o Apêndice J para obter uma lista de tipos de dados protegidos e definições.

Exemplos

Exemplo público 1

Tanto a Apple quanto a Samsung instituíram restrições ao uso do ChatGPT da OpenAI e do GitHub Copilot da Microsoft por alguns de seus funcionários devido a preocupações sobre o potencial de funcionários manipularem incorretamente e vazarem dados confidenciais da empresa. Essa medida se alinha a uma tendência crescente de empresas e governos em todo o mundo impondo restrições ao uso de plataformas GenAI. Em abril de 2023, a OpenAI lançou uma série de atualizações para o ChatGPT que permitiram melhores controles de privacidade depois que algumas nações expressaram suas preocupações.

Leia mais aqui: Apple restringe uso do ChatGPT por funcionários, Juntando-se a outras empresas cautelosas com vazamentos | WSJ

Exemplo público 2

A Samsung Electronics proibiu o uso de qualquer chatbot com tecnologia de IA e ChatGPT por seus funcionários devido a preocupações sobre vazamento de informações internas confidenciais. Esta decisão vem após um vazamento acidental de código-fonte confidencial por meio do ChatGPT, levando a empresa a emitir um memorando proibindo o uso de ferramentas GenAI. Embora a gravidade exata do vazamento seja desconhecida, os dados compartilhados com chatbots podem ser armazenados em servidores de propriedade de empresas externas que operam o serviço, como a OpenAI, sem que a Samsung tenha a capacidade de acessar ou excluir os dados.

Leia mais aqui: Samsung proíbe ChatGPT entre Funcionários após vazamento de código sensível | Forbes

Considerações importantes

ÿ As informações inseridas em um sistema GenAI podem se tornar parte de seu conjunto de dados de treinamento. Assim, quaisquer dados confidenciais inseridos como parte de um prompt podem ser usados em saídas para outros usuários do sistema, o que pode resultar em exposição não intencional ou uso indevido dessas informações. As informações confidenciais não devem ser usadas de nenhuma forma que possa levar ao compartilhamento das informações fora de seu uso pretendido ou autorizado.

ÿ A proveniência dos dados, a gestão dos direitos digitais e a compreensão dos direitos dos dados são essenciais para gerenciamento responsável de GenAI. Programas de gerenciamento de direitos digitais devem ser robustos para evitar uso não autorizado de dados.

ÿ Os sistemas GenAI podem armazenar dados e informações inseridas como prompts indefinidamente.

ÿ Jogadores adversários podem hackear o sistema para obter acesso a quaisquer dados confidenciais armazenados.

Melhores práticas

ÿ Não insira ou divulgue dados protegidos ou informações não públicas, como parte de um prompt ao usar uma ferramenta pública GenAI, a menos que você possa validar os direitos de uso dessa forma do originador. Consulte as versões mais recentes das políticas de segurança de informações do DOE para orientação específica. Procure orientação do departamento jurídico da sua organização.

ÿ Não confie na GenAI para gerar informações ou dados confidenciais ou de missão crítica, pois as informações usadas para treinar modelos de IA podem não ser precisas, completas ou sem viés. Revise e continue a aderir às políticas, procedimentos e guias existentes para garantir a conformidade com os requisitos de informação dos Laboratórios Nacionais e do DOE. Além disso, continue a seguir os requisitos existentes, como aqueles relacionados à qualidade, segurança da informação e integridade. Trabalhe com as PMEs e organizações de conformidade apropriadas do DOE e do Laboratório Nacional, como o Office of General Counsel, o Office of Export Control, o Classification Office, o Office of Environment, Health, Safety, and Security (EHSS) e outros, conforme apropriado.

ÿ Continue seguindo os procedimentos existentes de privacidade e segurança cibernética e revise iterativamente as políticas e procedimentos existentes para entender como a confidencialidade deve ser aplicada de forma geral dentro de uma organização, para garantir que novas políticas e regulamentações sejam implementadas nos procedimentos emitidos e para se manter atualizado com os requisitos mais recentes.

ÿ Incentivar o treinamento de engenharia rápida para usuários do GenAI para aprender as melhores maneiras de estruturar prompts que gerem resultados mais precisos (consulte o Apêndice I para obter mais detalhes sobre engenharia rápida). ÿ Revisar os resultados produzidos pelas ferramentas GenAI para garantir que quaisquer problemas relacionados à confidencialidade sejam resolvidos. identificados e abordados.

ÿ Praticar o armazenamento e processamento seguro de dados protegidos e informações não públicas e implementar controles de acesso. Defina claramente quem tem acesso aos dados e o propósito de seu uso todas as vezes.

ÿ Observe que as respostas à solicitação de contrato são informações proprietárias (confidenciais).

ÿ Consulte o treinamento existente do DOE sobre confidencialidade que os funcionários são obrigados a fazer.

ÿ É fundamental que os usuários entendam os direitos legais que o Governo (tem ou não tem) no dados inseridos em ferramentas de IA e/ou que são usados para treinar LLMs de dados.

Recursos adicionais

ÿ Treinamento obrigatório do DOE sobre confidencialidade: CUI-100DE Informações não classificadas controladas Visão geral

8.7 Propriedade Intelectual

As ferramentas GenAI introduzem riscos em torno da propriedade intelectual (PI), incluindo direitos autorais e proteções de dados, pois as ferramentas podem acessar trabalhos protegidos por direitos autorais e gerar saídas que se assemelham muito ao conteúdo desses trabalhos. Também deve ser observado que a interseção do GenAI e as leis, regulamentações e políticas existentes sobre PI e direitos autorais é dinâmica e evolutiva. Este guia não substitui o aconselhamento jurídico; portanto, quaisquer questões legais relacionadas à interseção do GenAI e PI devem ser direcionadas ao DOE ou ao consultor jurídico do contratado.

Definição

Propriedade intelectual (PI) é propriedade intangível que é o produto de um pensamento original, incluindo invenções, designs, escritos, imagens e nomes, muitos dos quais são protegíveis por direitos estatutários e contratuais, incluindo patentes, direitos autorais, segredos comerciais e marcas registradas (direitos de propriedade intelectual ou DPI). Dados também podem ser protegíveis como PI, normalmente como compilação com direitos autorais, se selecionados e organizados de forma única e original, como com um conjunto de dados. Dados em tal formato podem ser protegidos por direitos autorais e licenciáveis.

Direitos autorais não são um direito único, mas um conjunto de direitos que incluem não apenas a reprodução, mas também fornecem ao detentor dos direitos autorais o direito de impedir que outros adaptem, distribuam ao público, executem e exibam o trabalho protegido por direitos autorais (inclusive digitalmente).

Questões de propriedade intelectual têm implicações legais, financeiras e éticas significativas.

Exemplos

Exemplo público 1

A comediante Sarah Silverman e dois autores entraram com uma ação coletiva em 7 de julho de 2023 contra a OpenAI e a Meta, acusando-as de violação de direitos autorais pelo uso de seu trabalho protegido nos conjuntos de dados de treinamento das empresas. De acordo com o processo, “materiais protegidos por direitos autorais foram copiados e ingeridos como parte do treinamento”. Embora o resultado ainda esteja pendente, o contexto em torno deste processo é de grande importância, pois o conjunto de dados de treinamento pode incluir materiais protegidos por direitos autorais “sem permissão, raspando “bibliotecas paralelas” ilegais online que contêm o texto de milhares de livros”, conforme mencionado pelo The New York Times.

Leia mais aqui: Sarah Silverman processa OpenAI e Meta Sobre violação de direitos autorais | The New York Times

Exemplo público 2 Em

Andersen v. Stability AI et al., um caso aberto em 2022, três artistas entraram com uma ação judicial contra vários fornecedores da GenAI sob a alegação de que os sistemas GenAI usavam suas obras originais como parte de um conjunto de treinamento. Os usuários desses sistemas conseguiram gerar obras muito semelhantes às obras dos artistas originais. Se um tribunal decidir que as obras geradas pela GenAI são derivadas e não autorizadas, penalidades consideráveis podem ser aplicadas.

Leia mais aqui: A IA generativa tem um problema de propriedade intelectual | Harvard Business Review

Considerações importantes

ÿ De acordo com a lei atual, uma invenção criada exclusivamente por uma ferramenta de IA não pode ser patenteada ou protegido por direitos autorais porque a saída não foi criada por um humano. 30 Em agosto de 2023, o United States Copyright Office, uma subsidiária da Biblioteca do Congresso, divulgou uma solicitação de comentários públicos sobre a interação da IA e da lei de direitos autorais, que fornece uma imagem dos tipos de discussões que estão em andamento nesta área de estudo legal.31 Os inventores devem discutir casos específicos com um advogado de patentes, pois a política do Departamento de Energia não consegue determinar a legalidade. O US Copyright Office também emitiu várias declarações informando aos criadores que não registrará direitos autorais para obras produzidas por uma máquina ou programa de computador.

ÿ Em 13 de fevereiro de 2024, o Escritório de Patentes e Marcas dos Estados Unidos (USPTO) emitiu uma nova orientação que explica que, embora as invenções assistidas por IA não sejam categoricamente não patenteáveis, a análise de invenção deve se concentrar nas contribuições humanas, já que as patentes funcionam para incentivar e recompensar a engenhosidade humana. A proteção de patente pode ser buscada para invenções para as quais uma pessoa física forneceu uma contribuição significativa à invenção, e a orientação fornece procedimentos para determinando o mesmo. [32]

ÿ Em agosto de 2023, o United States Copyright Office, uma subsidiária da Biblioteca do Congresso, divulgou um pedido de comentário público sobre a interação da IA e da lei de direitos autorais, que fornece uma imagem dos tipos de discussões que estão em andamento nessa área de estudo jurídico. [33] Os inventores devem discutir casos específicos com um advogado de patentes, pois a política do Departamento de Energia não consegue determinar a legalidade. O US Copyright Office também emitiu várias declarações informando aos criadores que não registrará direitos autorais para obras produzidas por uma máquina ou programa de computador.

ÿ As ferramentas GenAI são treinadas em grandes conjuntos de dados raspados, e esse treinamento forma a base para as respostas do modelo aos prompts. Portanto, as ferramentas GenAI podem gerar saídas que contenham informações plagiadas ou protegidas por direitos autorais.

ÿ Deve-se tomar cuidado considerável para não infringir direitos de propriedade intelectual ou violar outras proteções ao inserir dados em prompts do GenAI ou usar dados para treinar LLMs.

Melhores práticas

ÿ Cumpra as políticas e procedimentos existentes relativos a questões de direitos autorais e continue monitorando mudanças nas leis de direitos autorais, políticas e procedimentos recomendados que se aplicam às ferramentas GenAI.

ÿ Tenha um humano no processo, de preferência alguém que tenha conhecimento do GenAI, para validar as fontes de saída geradas e evitar plágio e/ou problemas de direitos autorais.

ÿ Tenha cuidado ao usar as saídas de um modelo em outro trabalho e tenha em mente que a linguagem grande Os modelos não dirão com segurança se suas fontes são de domínio público ou não.

ÿ Quando as soluções GenAI desempenham um papel na criação de uma ideia, abordagem ou invenção no DOE ou em um Laboratório Nacional, os funcionários devem identificar claramente a contribuição específica (por exemplo, atribuição em um relatório, registro de laboratório, divulgação de invenção) e citar o GenAI como parte de sua metodologia de pesquisa. Pode ser essencial para determinar a patenteabilidade ou a capacidade de direitos autorais conhecer atribuições específicas a humanos versus tecnologias de IA. Vários guias de estilo e editoras estão desenvolvendo orientações sobre como creditar adequadamente as ferramentas de IA em trabalhos escritos (consulte os recursos adicionais abaixo). Os funcionários devem seguir essas diretrizes onde elas existirem.

ÿ Educar os usuários sobre o desafio da IA gerar conteúdo que pode infringir direitos autorais existentes e promover uma compreensão dos direitos de propriedade intelectual com relação à GenAI.

ÿ Use ferramentas secundárias para identificar e validar fontes, contexto e citações, especialmente nos casos em que o usuário tem *algum* conhecimento prévio.

ÿ Evite usar a saída do GenAI para criar conteúdo de site, a menos que a origem dos dados de treinamento seja verificada como apropriada para o uso fornecido, pois eles podem conter ou ter sido treinados em dados confidenciais ou sensíveis. Consulte a Seção 8.6: Confidencialidade.

ÿ Empregue as melhores práticas para engenharia rápida (consulte o Apêndice I para obter informações adicionais sobre engenharia rápida).

Recursos adicionais

ÿ O uso de materiais protegidos por direitos autorais por funcionários do governo, Departamento de Energia

ÿ Congressional Research Service: Inteligência Artificial Generativa e Lei de Direitos Autorais ÿ A IA generativa tem um problema de propriedade intelectual

ÿ “Como citar ChatGPT”, Timothy McAdoo, APA Style, 7 de abril de 2023

8.8 Segurança

Os sistemas GenAI devem ser projetados para serem seguros para os usuários do sistema e para a sociedade em geral. Os resultados dos sistemas GenAI não devem comprometer a segurança dos indivíduos ou sua saúde ou propriedade.

Definição

Os sistemas de IA são **seguros** se "não conduzirem, sob condições definidas, a um estado em que a vida humana, a saúde, a propriedade ou o ambiente estejam em perigo".34 A operação **segura** dos sistemas de IA é alcançada através de:

- ÿ Práticas responsáveis de design, desenvolvimento e implantação
- ÿ Informações claras aos implantadores sobre o uso responsável do sistema
- ÿ Tomada de decisão responsável por implantadores e usuários finais (por exemplo, por meio de aprendizagem por reforço, definido na Seção 5.1)
- ÿ Explicação e documentação de riscos com base em evidências empíricas de incidentes35

Exemplos

Exemplo público 1 Em 2022, foi relatado que veículos Tesla que utilizavam a funcionalidade Autopilot assistida por IA se envolveram em 273 acidentes durante o ano anterior. Isso incluiu acidentes com outros carros e motocicletas, e mortes de pedestres e motoristas.

A funcionalidade do piloto automático inclui a capacidade de manter a velocidade e a distância segura atrás de outros carros, permanecer dentro das linhas de suas faixas e fazer mudanças de faixa em rodovias. De acordo com a Tesla, no entanto, os motoristas humanos devem manter os olhos na estrada e as mãos no volante, com a tecnologia servindo como um assistente. Essa supervisão humana é crítica para a operação segura do veículo.

Leia mais aqui: Teslas com piloto automático envolvidos em 273 acidentes relatados desde o ano passado | The Washington Post

Exemplo público 2 O presidente Biden anunciou que as principais empresas de IA, como OpenAI, Alphabet e Meta, "fizeram compromissos voluntários com a Casa Branca para implementar medidas como marca d'água em conteúdo gerado por IA para ajudar a tornar a tecnologia mais segura". Isso permitirá a identificação de quando o conteúdo foi gerado por IA e, mais importante, a identificação de deepfakes que podem espalhar informações falsas ou ser usados para fraudar indivíduos. As empresas também se comprometeram a testar os sistemas completamente antes do lançamento e a se concentrar na proteção da privacidade dos usuários.

Esses compromissos são passos para garantir salvaguardas no GenAI.

Leia mais aqui: OpenAI, Google e outros prometem colocar marca d'água no conteúdo de IA para segurança, diz a Casa Branca | Reuters

Considerações importantes

- ÿ Vários tipos de riscos envolvendo segurança podem necessitar de estratégias personalizadas de mitigação de riscos de IA dependendo do contexto e da gravidade dos riscos potenciais.
- ÿ A segurança se relaciona principalmente ao uso e à aplicação do sistema. Riscos de segurança podem surgir tanto de negligência quanto de intenção deliberadamente maliciosa.
- ÿ Estabelecer mecanismos para apoiar a reprodutibilidade e a capacidade de examinar os resultados quanto à precisão e consistência por meio de controle de versão e procedência de entradas de treinamento, parâmetros de modelo, corpus de dados e outros elementos-chave do sistema.

Melhores práticas

ÿ Adote uma abordagem e mentalidade de segurança desde o projeto, considerando os riscos de segurança durante todo o ciclo de vida da IA, começando o mais cedo possível durante as fases de planejamento e design.

ÿ Não use o GenAI para atividades maliciosas ou enganosas, como a criação de malware, identidade roubo ou falsificação de identidade.

ÿ Desenvolver medidas de segurança para a implementação do sistema de IA, incluindo verificações contra ameaças nocivas e usos não intencionais.

ÿ Aproveitar as diretrizes de segurança nos setores de transporte e saúde e alinhar-se com as diretrizes ou padrões específicos do setor ou da aplicação existentes na estratégia de mitigação de riscos de segurança da IA (por exemplo, o NIST AI Risk Management Framework (NIST AI RMF)). ÿ Tenha um

humano no circuito durante todo o ciclo de vida da IA. Supervisão, validação e verificação humana, são um processo combinado e iterativo que começa na fase de planejamento e design e continua durante todo o ciclo de vida da IA, inclusive após a implantação do modelo.

ÿ Tenha os controles de soluções de IA responsáveis apropriados em vigor.

ÿ Considere usar o prompt (qualquer modalidade, como um sensor) como uma interface com o sistema sobre o estado atual que pode ser usado como um controle para aumentar a segurança.

ÿ Considere o uso secundário das saídas, garantindo que quaisquer ressalvas, considerações e/ou suposições sejam adicionadas à saída para que ela não seja inadvertidamente mal utilizada.
O uso malicioso intencional está fora do nosso controle.

ÿ De acordo com o NIST AI RMF, “as abordagens de gerenciamento de risco de segurança de IA devem seguir as orientações dos esforços e diretrizes para segurança em áreas como transporte e saúde e se alinhar com as diretrizes ou padrões existentes específicos do setor ou da aplicação”.36

8.9 Justiça e parcialidade

O GenAI apresenta desafios na definição, mensuração e abordagem de preocupações sobre justiça e preconceito de diversas maneiras.

Definição

A justiça na IA envolve abordar questões como preconceito prejudicial e discriminação para promover igualdade e equidade. Os padrões de justiça podem ser complexos e difíceis de definir porque as percepções de justiça diferem entre as culturas e podem mudar dependendo da aplicação.37

Os sistemas GenAI devem ser projetados para serem **justos ,** de modo que indivíduos ou grupos não sejam sistematicamente prejudicados por decisões orientadas por IA. Alcançar a justiça na IA pode ser desafiador, pois requer consideração cuidadosa de diferentes tipos de viés e usar a tecnologia de uma forma que evite favoritismo ou discriminação, particularmente para humanos.

Viés refere-se ao desvio sistemático e consistente da saída de um algoritmo do valor verdadeiro ou do que seria esperado na ausência de viés.38 Viés é um componente da justiça e vem em muitas formas, indo além da falta de equilíbrio demográfico ou representatividade de dados. O NIST identificou três categorias principais de **viés** de IA a serem consideradas e gerenciadas: *sistêmico, computacional e estatístico,* e *humano-cognitivo.* Cada um deles pode ocorrer na ausência de preconceito, parcialidade ou intenção discriminatória.

- *O viés sistêmico* pode estar presente em conjuntos de dados de IA, normas, práticas e processos organizacionais em todo o ciclo de vida da IA e na sociedade em geral que usa sistemas de IA.
- *Vieses computacionais e estatísticos* podem estar presentes em conjuntos de dados de IA e processos algorítmicos, e muitas vezes decorrem de erros sistemáticos devido a amostras não representativas.
- *Os preconceitos cognitivos humanos* estão relacionados com a forma como um indivíduo ou grupo utiliza as informações do sistema de IA para decidir ou preencher informações ausentes, ou como os humanos pensam sobre os propósitos e funções de um sistema de IA. Vieses cognitivos humanos são onipresentes em processos de tomada de decisão em todo o ciclo de vida da IA e uso do sistema, incluindo o design, implementação, operação e manutenção da IA.39

Embora justiça e preconceito sejam conceitos intimamente relacionados, eles diferem de maneiras importantes. A principal diferença é que, embora o preconceito possa ser não intencional, a justiça é inerentemente uma meta deliberada e intencional. Em outras palavras, o preconceito pode ser visto como uma questão técnica, enquanto a justiça é uma questão social e ética.40

Exemplos

Exemplo público 1

A solução GenAI de texto para imagem da Stable Diffusion foi identificada pela Bloomberg como um modelo que contribui para estereótipos raciais e de gênero tendenciosos. A Bloomberg usou a ferramenta da Stable Diffusion para criar milhares de imagens relacionadas a crime e emprego. Nesta análise, o modelo foi solicitado com texto para criar imagens de trabalhadores para 14 empregos — 300 imagens para sete empregos geralmente considerados como "bem pagos" nos EUA e 300 imagens para sete empregos geralmente considerados "mal pagos" — bem como três tópicos relacionados ao crime nos EUA. A análise descobriu que as imagens geradas para os empregos bem pagos eram de pessoas com tons de pele mais claros.

Em contraste, as imagens geradas com pele mais escura

Exemplo público 2 Uma

ação coletiva foi movida contra o fornecedor de software de gestão financeira e RH Workday, alegando que o software produziu um sistema de triagem que resultou em preconceito racial. A ação alega que a Workday "oferece ilegalmente um sistema de triagem de candidatos baseado em algoritmo que determina se um empregador deve aceitar ou rejeitar uma inscrição para emprego com base na raça, idade e/ou deficiência do indivíduo". O autor da ação afirma que as ferramentas de IA da Workday dependem de algoritmos que podem estar cheios de preconceito humano.

Leia mais aqui: Workday quer que alegação de algoritmo de recrutamento racialmente tendencioso seja rejeitada | The Register

Exemplos

os assuntos foram criados pela solução em resposta a prompts como “funcionário de fast-food”. A conclusão desta análise implica que a representação racial e de gênero em várias imagens de carreira era significativamente diferente da representação nas carreiras reais. Por exemplo, cerca de 3% das imagens geradas para o prompt “juiz” eram mulheres, enquanto na realidade, cerca de 34% dos juízes americanos são mulheres.

Leia mais aqui: Humanos são tendenciosos. IA generativa é ainda pior | Bloomberg

Considerações importantes

- ÿ A justiça precisa ser definida para cada caso de uso no início da fase de design e planejamento, pois pode significar coisas diferentes em contextos diferentes.
- ÿ Os resultados dos sistemas GenAI podem e irão produzir viés se o viés for incluído nos conjuntos de dados de treinamento, validação ou teste, o que geralmente é o caso.
- ÿ O viés pode ser introduzido em qualquer ponto do ciclo de vida da IA.
- ÿ Dados não estacionários, ou dados que mudam de forma dinâmica e imprevisível ao longo do tempo, podem produzir viés.
- ÿ A escala de viés e/ou informações não factuais sai do controle mais rapidamente com o GenAI do que com IA não generativa.
- ÿ Quando os resultados são injustos ou incluem preconceitos, as decisões que eles informam podem impactar as comunidades que podem ser afetadas por esses algoritmos de forma injusta e tendenciosa.

Melhores práticas

- ÿ Não confie somente nos sistemas GenAI para tomar decisões. Em vez disso, use os sistemas para ajudar a informar decisões.
- ÿ Incorpore justiça algorítmica por design, incluindo conceitos de justiça em todo o ciclo de vida da IA. Considere desenvolver e usar uma lista de verificação de justiça e viés para cada estágio.
- ÿ Implementar uma gestão de dados equitativa para considerar e priorizar cuidadosamente as necessidades das comunidades desfavorecidas quando se trata de gestão de dados e informações.
- ÿ Garantir que existam estratégias centradas no ser humano durante todo o ciclo de vida da IA que possam ser usadas no caso de um sistema GenAI com falha que afete os direitos das pessoas, inclusive tendo sempre um humano no circuito para verificar se os resultados são representativos e não terão consequências negativas que possam afetar considerações de justiça ou preconceito.
- ÿ Tenha em mente a intenção inicial por trás da existência do sistema durante todo o ciclo de vida da IA e verifique regularmente se o sistema está no caminho certo para o uso e os resultados pretendidos.
- ÿ Garantir que os sistemas de IA operem de forma justa e transparente, estabelecendo métricas de equidade e revisar regularmente os sistemas GenAI e a saída para conformidade e representação. Entender e mitigar os riscos de resultados discriminatórios para manter acesso equitativo e se beneficiar do GenAI.
- ÿ Desenvolver procedimentos para revisão, compartilhamento e avaliação de sistemas GenAI através da lente de imparcialidade e viés. Isso inclui realizar análises nos estágios iniciais da iniciativa GenAI para se tornar ciente dos diferentes vieses que podem ocorrer, bem como implementar procedimentos que identifiquem e mitiguem o viés em dados de treinamento e que usem conjuntos de dados diversos para treinamento para evitar amplificar o viés existente.

- Treinar, validar e testar modelos GenAI usando conjuntos de dados representativos que sejam tão "justos" quanto possível, conforme definido pelo contexto do caso de uso no início da fase de design do projeto. Isso inclui garantir que o modelo de treinamento seja representativo da sociedade em geral ou da sociedade dado o tópico em questão e avaliar a justiça em conjuntos de dados identificando representação, limitações correspondentes e quaisquer correlações prejudiciais ou discriminatórias entre recursos, rótulos e grupos.
- Após tentar abordar a justiça no conjunto de dados, incluindo filtragem se necessário, verifique novamente as saídas do modelo para ver se algum *novo* viés artificial foi criado. Aplique ajustes de normalização ou outros equilíbrios matemáticos para corrigir novos vieses não intencionais ou distorções da justiça percebida introduzidas por tentativas anteriores de abordar vieses originais.
- Verifique os sistemas de IA quanto a vieses injustos e considere os efeitos dos vieses criados por saídas anteriores incorporadas aos conjuntos de treinamento e os loops de feedback que isso pode criar.
- Teste os resultados que o modelo GenAI produz fazendo perguntas com respostas conhecidas. Dê exemplos de perguntas feitas de maneiras diferentes para testar como os resultados mudam.
- Forneça advertências e suposições com relação às fontes GenAI na saída, bem como notas de rodapé.
- Tenha cuidado extra quando o GenAI for usado para uma atividade que pode ser regulamentada pela Lei de Igualdade de Oportunidades de Emprego, Lei dos Direitos Civis ou Lei dos Americanos com Deficiências (por exemplo, contratação, triagem de currículos, recrutamento ou qualquer outra função de RH).
- Fique de olho em ferramentas, processos e organizações que podem verificar se há IA confiável (para incluir detecção de viés). Este é um mercado emergente onde soluções valiosas podem surgir.
- Esteja ciente do mercado e da reputação das ferramentas em uso e do fato de que esses fatores podem e mudará rapidamente.
- Desenvolver continuamente a conscientização em toda a organização sobre justiça e identificação de preconceitos, inclusive educando os usuários sobre como os preconceitos nos dados podem levar a resultados tendenciosos.

Recursos adicionais

- Plano de Consistência DOE EO 13960
- A IA generativa leva estereótipos e preconceitos de mal a pior
- AI Justiça 360

8.10 Alucinações e interpretações errôneas da IA

Alucinações ocorrem quando um sistema GenAI produz resultados que não são baseados em dados reais ou existentes, mas, em vez disso, são frequentemente conteúdos imaginativos ou irrealistas gerados além do conjunto de treinamento do sistema. Essas alucinações podem levar à disseminação de informações falsas.

Definição

Uma **alucinação de IA** ocorre quando um sistema GenAI fornece uma resposta que inclui informações irrelevantes, falsas ou sem sentido. Essas respostas são frequentemente articuladas e confiantes, mas contêm informações que não são verdadeiras. Observe que quando o modelo alucina, ele não está mentindo intencionalmente, pois não é motivado a enganar os usuários, nem tem consciência de que está fornecendo informações falsas. Alucinações também são conhecidas como erros gerados, confabulações, delírios ou fabricações. Elas ocorrem quando o sistema de IA interpreta mal seus dados de treinamento e os usa para criar respostas que não são factuais.

Tenha em mente os conceitos de validação, confiabilidade e explicabilidade ao discutir alucinações e as melhores práticas para mitigá-las.

- **Validação** é a “confirmação, através do fornecimento de evidências objetivas, de que os requisitos para um uso ou aplicação específica pretendida foram cumpridos”.41
- **A confiabilidade** é definida no mesmo padrão como a “capacidade de um item de funcionar conforme necessário, sem falhas, por um determinado intervalo de tempo, sob determinadas condições”.42
- **Explicabilidade** refere-se a uma representação dos mecanismos subjacentes à operação dos sistemas de IA, enquanto interpretabilidade refere-se ao significado da saída dos sistemas de IA no contexto dos seus propósitos funcionais concebidos.43

Exemplos

Exemplo público 1

Durante os testes do mecanismo de busca Bing, da Microsoft, com tecnologia de IA, os usuários vivenciaram um comportamento perturbador e ofensivo do chatbot, que se envolveu em conversas hostis e abusivas. Incidentes semelhantes ocorreram com outros chatbots, levantando preocupações sobre o desenvolvimento e a implantação responsáveis de ferramentas de IA. Ficou clara a pressa da Microsoft em lançar um chatbot com tecnologia de IA sem conduzir estudos completos sobre seu potencial de respostas inapropriadas durante interações prolongadas do usuário. Testes extensivos poderiam ter ajudado a identificar esses problemas. Enquanto as empresas trabalham para refinar esses sistemas e implementar limites nas interações do usuário, os incidentes destacam os desafios de gerenciar a saída de chatbots de IA e a necessidade de pesquisa e desenvolvimento contínuos neste campo.

Leia mais aqui: O novo chatbot de IA da Microsoft tem dito algumas coisas 'loucas e desequilibradas' | NPR

Exemplo público 2 Um advogado envolvido em um processo de danos pessoais contra uma companhia aérea usou o chatbot de IA ChatGPT para preparar um processo, mas acabou apresentando casos falsos ao tribunal. O advogado, Steven Schwartz, alegou que não sabia que a ferramenta, que ele confundiu com um mecanismo de busca, geraria informações falsas. O juiz está considerando impor sanções, pois este caso destaca a questão das alucinações de IA e levanta questões sobre o tratamento da comunidade jurídica de conteúdo gerado por IA. O incidente gerou preocupações sobre a capacidade de detectar falsificações geradas por IA e o impacto potencial na confiança na sociedade. As preocupações são altas de que a GenAI está avançando além da capacidade humana de detectar falsificações, o que levará a uma maior falta de confiança na sociedade.

Leia mais aqui: Advogado usou ChatGPT em tribunal e citou casos falsos. Juiz considera sanções | Forbes

Considerações importantes

ÿ Alucinações ocorrem frequentemente com sistemas GenAI e podem ser difíceis de identificar quando ocorrem em uma resposta geralmente coerente e articulada.

ÿ Sem nenhuma maneira de verificar as informações, as respostas alucinadas podem ser interpretadas pelo usuário como precisas e factuais.

ÿ Os conjuntos de dados de treinamento, especialmente aqueles que compreendem dados desconhecidos, podem incluir dados não factuais ou informações sem sentido que podem contribuir diretamente para alucinações.

ÿ A falta de transparência no conjunto de treinamento, algoritmo e modelo cria um efeito de “caixa preta”, onde não há percepção de como o sistema GenAI chegou à sua saída alucinatória imprecisa ou não factual.

ÿ Quando as alucinações ocorrem com frequência, o modelo GenAI não está alinhado com os princípios de validação, confiabilidade e explicabilidade.

Melhores práticas

ÿ Tenha um humano no circuito para verificar a precisão e a validade dos resultados. ÿ

Desenvolver a alfabetização em IA entre os usuários para que eles entendam as limitações e os riscos potenciais do GenAI. Concentre-se em ensinar como gerenciar prompts adequadamente e direcionar a saída desejada.

ÿ Use as melhores práticas de engenharia rápida para mitigar o risco de alucinações (consulte o Prompt Informações de engenharia no Apêndice I).

ÿ Use aprendizado por reforço com feedback humano (RLHF) para melhorar a confiabilidade e a precisão do modelo e aprimorar o alinhamento do modelo com seus usuários humanos. RLHF envolve uma equipe de humanos que pontuam as saídas do modelo (por exemplo, pontuações altas para saídas confiáveis e factuais e pontuações baixas para saídas alucinadas), e o modelo é subsequentemente treinado para gerar saídas que são mais aceitáveis para a equipe de humanos.44

ÿ Empregue “grounding” como uma técnica para mitigar o risco de alucinações. Use informações externas de fontes confiáveis para solicitar que o modelo GenAI gere uma resposta com base nas informações recuperadas e factuais dentro do contexto fornecido. A técnica de aterramento mais comum usada atualmente com LLMs é a “Retrieval Augmented Generation (RAG). Usando essa abordagem, a entrada para o LLM é alavancada para recuperar as informações de aterramento de um banco de dados especificado e, em seguida, alimentá-las ao LLM junto com a entrada original do usuário. Se o modelo não tiver a capacidade de responder com base apenas nas informações relevantes usadas para aterrar o sistema E dentro do contexto apropriado, o modelo retornará “informações insuficientes” em vez de uma resposta alucinada. 45

Recursos adicionais

ÿ Curso de Engenharia Prompt: ChatGPT Engenharia Prompt para Desenvolvedores, OpenAI

ÿ Compreendendo alucinações em IA: um guia abrangente, Pinecone

8.11 Lista de verificação de melhores práticas

Esta seção fornece um resumo das melhores práticas do GenAI, divididas em três categorias: **Pessoas, Organização** e **Tecnologia.** A categoria Pessoas contém as melhores práticas relacionadas à equipe, liderança, especialistas no assunto (SMEs) e desenvolvimento da força de trabalho. A categoria Organização inclui as melhores práticas relacionadas a processos, políticas, necessidades de negócios e governança. A categoria Tecnologia contém as melhores práticas em desenvolvimento técnico, implementação, monitoramento e segurança de IA. Cada uma das três categorias inclui conteúdo relacionado às melhores práticas específicas de função e às melhores práticas de gerenciamento de dados. Dentro de cada categoria, as melhores práticas são organizadas por sua localização relativa dentro do ciclo de vida da IA. Para obter informações adicionais sobre o Ciclo de Vida da IA, consulte o Apêndice G.

De acordo com a Ordem Executiva 14110 sobre o *Desenvolvimento e Uso Seguro, Confiável e Confiável de Inteligência Artificial,* uma variedade de novas políticas, procedimentos, padrões e melhores práticas serão criadas em relação ao desenvolvimento e uso de IA no espaço federal, incluindo "diretrizes e limitações... sobre o uso apropriado de GenAI". 46 Consulte esses recursos à medida que forem desenvolvidos e publicados.
A Ordem Executiva 14110 também incentiva o desenvolvimento de novos bancos de testes de capacidade de IA e ferramentas de avaliação de modelos, facilitados pelo Departamento de Energia em colaboração com outras partes interessadas. Esses bancos de testes e capacidades de avaliação podem fornecer suporte crítico na adesão às melhores práticas detalhadas abaixo.47

8.11.1 Melhores práticas gerais: Pessoas

- ÿ Consultar um grupo diversificado de especialistas no assunto (PMEs), partes interessadas e comunidades impactadas desde o início do projeto para identificar os benefícios e riscos do sistema, para ajudar a moldar a solução, aconselhar sobre o uso e, se apropriado, realizar uma avaliação liderada por um terceiro independente ou por especialistas que não atuem como desenvolvedores principais do sistema.48 O processo de teste de tecnologias GenAI deve ser colaborativo em todo o DOE
- ÿ Tenha um humano no circuito durante todo o ciclo de vida da IA. Aplique um processo iterativo começando na fase de planejamento e design e continuando durante todo o ciclo de vida da IA para verificar a precisão, validade, justiça, representatividade, viés, plágio e/ou problemas de direitos autorais. Desenvolva procedimentos específicos para garantir que essas verificações ocorram.
- ÿ Sistemas GenAI piloto, tornando os líderes empresariais e engenheiros participantes ativos na processo de pilotagem. Os ciclos de pilotagem são rápidos e focam em experimentos curtos e casos de uso que fornecem valor estratégico ao mesmo tempo em que mitigam riscos e eliminam casos de uso que não são viáveis. 49
- ÿ Treinar os usuários para entender os riscos de segurança associados à IA, bem como questões de preconceito e imparcialidade. Incluir treinamentos relevantes existentes do DOE sobre esses tópicos nos currículos de IA, quando disponíveis.
- ÿ Educar os usuários sobre o desafio da IA gerar conteúdo que pode infringir direitos autorais ou propriedade intelectual existentes e observar as contribuições da IA para a criação de uma ideia, abordagem ou invenção no DOE ou no Laboratório Nacional em metodologia de pesquisa.
- ÿ Atualizar descrições de cargos, práticas de recrutamento e documentos de desenvolvimento da força de trabalho para garantir que a expertise técnica específica do GenAI esteja sendo priorizada em responsabilidades de trabalho, qualificações, treinamentos, orientações e recursos relevantes, tendo em mente os casos de uso priorizados atuais do DOE (consulte o Inventário de Casos de Uso do DOE).
- ÿ Fornecer treinamentos atualizados sobre mitigação de riscos, engenharia rápida e melhores práticas de uso do GenAI para o pessoal envolvido em todo o ciclo de vida da IA. = Incluir treinamentos personalizados de privacidade do GenAI para funcionários com acesso a dados protegidos ou não públicos.
- ÿ Proporcionar oportunidades de qualificação dos funcionários existentes — formações, materiais de aprendizagem relevantes, programas de alfabetização digital, etc. — e atualizar os materiais regularmente.
- ÿ Fornecer recursos para educar os usuários gerais da organização sobre como as ferramentas GenAI não necessariamente fornecer resultados factuais. Os usuários devem entender que os resultados das ferramentas GenAI devem

ser considerado um “primeiro rascunho”. Uma opção para denotar esses documentos é a inclusão de marcas d’água marcando o documento como um rascunho inicial ou saída de IA.

8.11.2 Melhores práticas gerais: organização

- ÿ Ao começar a planejar qualquer iniciativa GenAI, defina os objetivos do sistema GenAI e os objetivos pretendidos. propósito. Os limites de escopo de dados e casos de uso devem ser documentados explicitamente.
- ÿ Tenha em mente a intenção e o propósito inicial do sistema e verifique regularmente o uso e os resultados alinhar-se a essa intenção.
- ÿ Adote uma abordagem e mentalidade de segurança desde o projeto e justiça desde o projeto, considerando os riscos em todo o ciclo de vida da IA e incluindo verificações desses princípios no sistema.
- ÿ Estabelecer um plano para identificar e mitigar rotineiramente os riscos e vulnerabilidades que surgem de um sistema GenAI, incluindo aqueles relatados pelos usuários.50
- ÿ Avaliar a imparcialidade e a representatividade em conjuntos de dados identificando a representação, limitações e quaisquer correlações prejudiciais ou discriminatórias entre características, rótulos e grupos.
- ÿ Empregar uma abordagem ágil nas comunicações organizacionais sobre tópicos e questões relacionadas à IA, como elas mudarão com a rápida evolução da tecnologia.
- ÿ Implementar a administração completa do ciclo de vida, que é a prática de proteger e proteger dados, metadados e informações ao longo de seu ciclo de vida. Isso inclui coleta, armazenamento, uso, controle, processamento, publicação, transferência, retenção e disposição. Documente esse processo completamente.
- ÿ Desenvolver ferramentas de medição e documentação de dados para capturar informações sobre o sistema GenAI e seu treinamento, operação e saídas no ambiente de produção. 51
- ÿ Verifique os sistemas GenAI e seus conjuntos de dados de treinamento quanto a vieses injustos. Considere os efeitos dos vieses criados por saídas anteriores incorporadas aos conjuntos de treinamento e os loops de feedback que isso pode criar.52 Implemente procedimentos que identifiquem e mitiguem esses vieses e use dados de treinamento diversos sempre que possível.
- ÿ Revise e continue a aderir às políticas, procedimentos e guias novos e existentes para garantir conformidade com os requisitos de informação do Laboratório Nacional e do DOE, inclusive quando se trata de tópicos como qualidade, privacidade, PI, segurança da informação e integridade.
- ÿ Estabelecer um plano para identificar e mitigar rotineiramente os riscos e vulnerabilidades que surgem dos sistemas GenAI, incluindo aqueles relatados pelos usuários.53
- ÿ Crie planos de mitigação para prováveis vetores de ataque de IA, como deepfakes e envenenamento de dados (quando um jogador adversário polui dados de treinamento para manipular a saída do modelo) e tenha em mente que nenhuma defesa infalível existe atualmente para prevenir completamente tais ataques. Consulte a publicação do NIST “Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations” para obter mais informações.54
- ÿ Exigir que fornecedores e desenvolvedores que projetem sistemas GenAI e criem recursos com recursos humanos interfaces voltadas para fornecer feedback rastreável sobre o status do sistema e procedimentos claros para operadores treinados para ativar e desativar funções do sistema.55
- ÿ Rever contratos com fornecedores terceirizados e acordos relativos ao compartilhamento de informações específicas de IA riscos.
- ÿ Garanta que a liderança sênior tenha revisado os riscos do sistema GenAI. A liderança precisa estar ciente de que ela é, em última análise, responsável pelo uso e função adequados do sistema GenAI.
- ÿ Considere formar e apoiar estruturas de governança fortes como parte do sistema GenAI implementações, incluindo governança formal de dados e iniciativas de governança de modelos com conselhos de governança organizacional alinhados.

8.11.3 Melhores práticas gerais: Tecnologia

- ÿ Comece por fornecer um espaço seguro (por exemplo, uma caixa de areia) para experimentação e dimensionamento quando o risco residual for considerado aceitável.56
- ÿ Pratique as melhores práticas de minimização de dados ao coletar dados para um sistema.
- ÿ Considere usar dados sintéticos (gerados algoritmicamente e usados como substitutos de dados reais) para mitigar riscos relacionados à privacidade, integridade de dados e dados insuficientes.57 Outra opção seria utilizar dados disponíveis publicamente do Data.gov ou relatórios públicos do DOE, se aplicável e legalmente apropriado.
- ÿ Garantir a observabilidade do sistema (ou seja, a extensão em que os estados internos de um sistema podem ser inferidos a partir de dados disponíveis externamente) por meio de três pilares de dados de observabilidade: métricas, logs e rastros.58
- ÿ Incentivar as melhores práticas para engenharia rápida para garantir precisão e eficácia.
- ÿ Não inclua informações não públicas (classificadas, privadas, sensíveis, confidenciais, etc.) como parte de um prompt/entrada para qualquer sistema GenAI comercial ou aberto. Consulte o Apêndice J para obter exemplos de dados protegidos.
- ÿ Não use o GenAI para atividades maliciosas ou enganosas, como criação de malware, roubo de identidade ou representação falsa de identidade.
- ÿ Estabelecer mecanismos para dar suporte à reprodutibilidade e à capacidade de analisar saídas quanto à precisão e consistência por meio do controle de versão e procedência de entradas de treinamento, parâmetros de modelo, corpus de dados e outros elementos-chave do sistema.
- ÿ Desenvolver sistemas robustos e seguros e capacidades de detecção para sites, atualizando-os regularmente para defender-se contra ameaças e planejar a resiliência do sistema.
- ÿ Estabelecer um programa para conduzir testes adversários ou red teaming.59

9. Conclusão

GenAI é uma tecnologia emergente com um grande potencial para impulsionar inovação, eficiência e valor no DOE. A adoção do GenAI está explodindo em todo o mundo, e a variedade de ferramentas disponíveis no mercado continua a se expandir. A tecnologia GenAI pode fornecer valor ao executar certas tarefas de rotina, como resumir um documento ou redigir um e-mail, em um período de tempo substancialmente menor do que os funcionários humanos. O futuro do local de trabalho do DOE pode ser aquele em que existe uma relação simbiótica entre funcionários humanos e ferramentas GenAI.

A tecnologia GenAI está evoluindo rapidamente, e o panorama de riscos e considerações ainda não é completamente compreendido. O ritmo em que o GenAI está avançando significa que o DOE precisa permanecer ágil e atualizado sobre o pensamento atual. Este documento descreve casos de uso em potencial, riscos proeminentes e melhores práticas para uso do GenAI no contexto do DOE. Espere que esses tópicos continuem a evoluir à medida que surgem casos de uso e riscos adicionais, à medida que as melhores práticas são refinadas e à medida que legislações e regulamentações adicionais são publicadas pelo Governo dos EUA. Este documento será atualizado regularmente para refletir o panorama em evolução do GenAI.

10. Apêndices

Apêndice A. Agradecimentos

O DOE Generative AI Reference Guide foi desenvolvido pelo Department of Energy Office of the Chief Information Officer (OCIO) com o apoio da Ernst & Young LLP. Os autores deste Reference Guide e o Department of Energy OCIO gostariam de agradecer à DOE Generative AI Reference Guide Tiger Team, que incluiu representantes das seguintes organizações, por contribuir com seu tempo e experiência:

- ÿ Administração de energia de Bonneville **(BPA)**
- ÿ Laboratório Nacional de Brookhaven **(BNL)**
- ÿ Tecnologias Críticas e Emergentes **(TEC)**
- ÿ Eficiência Energética e Energias Renováveis **(EE)**
- ÿ Meio Ambiente, Saúde, Segurança e Proteção **(EHSS)**
- ÿ Gestão Ambiental **(GA)**
- ÿ Energia Fóssil e Gestão de Carbono **(FECM)**
- ÿ Laboratório Nacional de Idaho **(INL)**
- ÿ Laboratório Nacional Lawrence Berkeley **(LBNL)**
- ÿ Administração Nacional de Segurança Nuclear **(NNSA)**
- ÿ Laboratório Nacional de Energias Renováveis **(NREL)**
- ÿ Laboratório Nacional de Oak Ridge **(ORNL)**
- ÿ Gabinete de Ciência **(SC)**
- ÿ Laboratório Nacional do Noroeste do Pacífico **(PNNL)**
- ÿ Laboratórios Nacionais de Sandia **(SNL)**
- ÿ Laboratório Nacional de Savannah River **(SRNL)**
- ÿ Sítio do Rio Savannah **(SRS)**
- ÿ Laboratório Nacional de Aceleradores de Stanford **(SLAC)**
- ÿ Acelerador Nacional Thomas Jefferson (Laboratório de Engenharia Civil)
- ÿ Subsecretário de Ciência e Inovação **(S4)**
- ÿ Administração de energia da área ocidental **(APA)**

Além disso, gostaríamos de reconhecer os seguintes colaboradores por suas contribuições significativas e feedback valioso, incluindo os copresidentes da equipe Tiger, a equipe patrocinadora e indivíduos da equipe Tiger.
Equipe que se destacou em sua participação:

Arão Haglund	Gardy Rosius (Copresidente)	Lance Roeske
Ahmad Sultão	Greg Doan	Malaquias Schram
Brad Wilson	James P. Vivo	Margaret Lentz
Brian Post	Jason Talley	Maria McClelland
Bridget Carper (Executiva Patrocinador)	Jayu Wu	Randy Boi
Brooke Dickson	Jodi Kouts (Patrocinador)	Robert Rei
Cristão Stauffer	Jonnie Bradley (Copresidente)	Rochelle Blaustein
Darrell Beschen	Kathleen Oprea	Sandra Logan (Patrocinadora)
Elaine Ulrich	Ken Caça	Steven Wong
Érica Vosseller	Kenneth Calabrese (patrocinador) Tom Harper (copresidente)	
Félix González	Kerstin Kleese Van Dam	Vicki Michetti (Executiva Patrocinador)

Por fim, gostaríamos de agradecer à Diretora de Informações do Departamento de Energia, Ann Dunkin, por sua liderança e apoio durante todo esse processo.

Apêndice B. Documentos Referenciados

Referências internas do DOE

- ÿ Plano de Consistência DOE 13960
- ÿ DOE CUI Slicksheet
- ÿ Início — Diretrizes, Orientações e Delegações do DOE
- ÿ Padrões de Conduta Ética - Memorando de Política nº 93
- ÿ Ficha de Informação Não Classificada Controlada

Referências Públicas

- ÿ Manual de gerenciamento de risco de IA do DOE (AIRMP)
- ÿ Inventário de casos de uso de IA do DOE 2023
- ÿ Regulamento Federal de Aquisições (FAR)
- ÿ "Programa de Informação de Infraestrutura Crítica Protegida (PCII)", Segurança Cibernética e Agência de Segurança de Infraestrutura (CISA)
- ÿ O que é Propriedade Intelectual? ÿ
- Congressional Research Service: Inteligência Artificial Generativa e Lei de Direitos Autorais ÿ A IA generativa tem um problema de propriedade intelectual

ÿ O que são direitos autorais?

ÿ A IA generativa leva estereótipos e preconceitos de mal a pior

ÿ AI Justiça 360

Apêndice C. Recursos de aprendizagem externos relevantes

ÿ Engenharia Rápida

- o ChatGPT Prompt Engineering para desenvolvedores, DeepLearning.AI ChatGPT 101: Potencialize seu trabalho e sua vida (mais de 750 prompts incluídos)

ÿ ChatGPT Masters: IA generativa, engenharia de prompts, Chat GPT | Udemy ÿ Aprendizagem do Google GenAI

- o Novos recursos de treinamento de IA generativa do Google Cloud: Sete novos treinamentos gratuitos de GenAI Cursos | Blog do Google Cloud

ÿ Aprendizagem de IA aberta da Microsoft

- o Uma introdução ao serviço Azure OpenAI o Explore como os clientes estão colocando o Azure AI para trabalhar para eles o Azure OpenAI: Briefing de negócios o Azure OpenAI: Briefing técnico
- o Azure OpenAI: Saiba mais sobre ChatGPT e DALL·E
- o Página do produto Azure OpenAI

ÿ ChatGPT, LLMs e IA generativa: o que sua empresa precisa saber

ÿ IA generativa para marketing

ÿ Adotando a IA: O futuro da criação de conteúdo, Ernesto Anaya, TEDxSCCS Youth

ÿ Como RH e TI podem usar IA para construir uma força de trabalho baseada em habilidades

ÿ Webinar - Papel da IA Generativa no Futuro do Recrutamento

ÿ Curso Online Gratuito Elementos de IA

Apêndice D. Treinamentos e recursos de aprendizagem relevantes do DOE (localizados no Núcleo de Aprendizagem do DOE)

ÿ Introdução à Inteligência Artificial (ID do Learning Nucleus: 55476) ÿ AI-900: Fundamentos do Azure AI: IA e ML (ID do Learning Nucleus: 84436) ÿ Inteligência Artificial: Teoria Básica de IA (ID do Learning Nucleus: 76001) ÿ Inteligência Artificial: Visão Geral da Interação Homem-computador (ID do Learning Nucleus: 79429) ÿ Inteligência Artificial: Metodologias de Interação Homem-computador (ID do Learning Nucleus: 79430) ÿ Inteligência Artificial: Tipos de Inteligência Artificial (ID do Learning Nucleus: 76002) ÿ Elementos de um Arquiteto de Inteligência Artificial (ID do Learning Nucleus: 79433)

- ÿ AI-900: Fundamentos do Azure AI: Usando o Azure Machine Learning Studio (ID do Learning Nucleus: 84438)
- ÿ AWS Certified Machine Learning: Serviços de IA/ML (ID do núcleo de aprendizagem: 84340)
- ÿ Visão geral de informações não classificadas controladas CUI-100DE

Apêndice E. Mais detalhes sobre políticas, diretrizes e referências federais

1. Memorando do Escritório de Gestão e Orçamento, Promovendo Governança, Inovação e Gestão de Riscos para o Uso de Inteligência Artificial por Agências, 28 de março de 2024

Este memorando aborda um subconjunto de riscos de IA, bem como questões governamentais e de inovação diretamente vinculadas ao uso de IA por agências. Os riscos abordados neste memorando resultam de qualquer dependência de resultados de IA para informar, influenciar, decidir ou executar decisões ou ações da agência, o que pode prejudicar a eficácia, segurança, equidade, justiça, transparência, responsabilidade, adequação ou legalidade de tais decisões ou ações. Consistente com a Seção 104(c) e (d) da Lei de IA no Governo de 2020, dentro de 180 dias da emissão deste memorando, e a cada dois anos depois disso, o DOE emitirá um plano para obter consistência com este memorando. As agências também devem incluir planos para atualizar quaisquer princípios e diretrizes internas de IA existentes para garantir a consistência com este memorando.

2. Ordem Executiva 14110: Desenvolvimento e uso seguro, protegido e confiável de inteligência artificial, outubro de 2023

A Ordem Executiva 14110 descreve um novo conjunto de definições para a terminologia-chave de IA, ao mesmo tempo em que descreve uma variedade de novas ações a serem tomadas pelo Poder Executivo, agências federais, instituições de pesquisa, o setor privado e parceiros internacionais para aumentar o uso de IA, ao mesmo tempo em que gerencia o risco relacionado e protege os direitos civis e os direitos dos trabalhadores. A EO também determina a criação de Chief AI Officers (CAIOs) em agências federais que atendam aos critérios exigidos e busca acelerar a contratação e o treinamento de profissionais de IA em todo o governo.

3. Congressional Research Service (CRS), Generative Artificial Intelligence and Data Privacy: A Primer, maio de 2023

A IA generativa apresenta riscos em termos de privacidade, desinformação, violação de direitos autorais e potencial geração de imagens sexuais não consensuais devido a fontes de dados de treinamento. Divulgação clara e consentimento afirmativo, particularmente em áreas sensíveis como assistência médica ou serviços jurídicos, são essenciais. Embora os EUA não tenham leis abrangentes de privacidade de dados, regulamentações estaduais específicas, como a Lei de Proteção à Privacidade Online de Crianças (COPPA), podem se aplicar a alguns aplicativos GenAI. Discussões contínuas e potencial legislação federal são necessárias para abordar as implicações de privacidade do GenAI e do uso de dados.

4. Inteligência Artificial Generativa e Direito Autoral, maio de 2023

Decidir sobre a propriedade de direitos autorais para conteúdo gerado por programas de IA pode ser complexo. O US Copyright Office atualmente reconhece direitos autorais em obras criadas por seres humanos, mas há debates e litígios em andamento sobre se obras geradas por IA podem ser elegíveis para proteção de direitos autorais. Enquanto a questão permanece incerta, os tribunais podem considerar fatores como envolvimento humano, arranjos criativos ou modificações e termos contratuais para determinar a propriedade de direitos autorais em obras geradas por IA.

5. Relatório do Ano 1 do Comitê Consultivo Nacional de Inteligência Artificial (NAIAC), maio de 2023

O National Artificial Intelligence Advisory Committee (NAIAC) aconselha o Presidente sobre o impacto da IA em várias áreas. Este relatório compila as descobertas do comitê, incluindo temas de alto nível, objetivos,

ações e um plano para futuras atividades do comitê. O relatório está organizado em quatro temas: (1) Liderança em Inteligência Artificial Confiável, (2) Liderança em Pesquisa e Desenvolvimento, (3) Suporte à Força de Trabalho dos EUA e Criação de Oportunidades, e (4) Cooperação Internacional. O relatório enfatiza que a IA é uma tecnologia que exige atenção imediata, substancial e sustentada do governo.

6. Estrutura de gerenciamento de risco de IA do NIST (NIST AI RMF 1.0), janeiro de 2023

O NIST AI RMF auxilia usuários e desenvolvedores na análise e mitigação de riscos de IA com diretrizes práticas e melhores práticas para sua implementação. A estrutura tem duas partes: enquadramento de risco e funções principais (governar, mapear, medir e gerenciar). Ele fornece sete considerações de risco e confiança para IA, particularmente para ferramentas GenAI, promovendo desenvolvimento e implementação responsáveis e confiáveis. Para obter detalhes adicionais sobre o NIST AI RMF 1.0, consulte o Apêndice H.

7. Lei de promoção da IA americana, dezembro de 2023

O Advancing American AI Act foi originalmente introduzido em abril de 2021 e eventualmente aprovado como parte do James M. Inhofe National Defense Authorization Act para o ano fiscal de 2023. O objetivo deste ato é cumprir as missões da agência por meio do "uso de tecnologias inovadoras de inteligência artificial aplicada". 60 O projeto de lei exige que as agências tomem medidas para promover o uso da IA, ao mesmo tempo em que tomam medidas para mitigar os riscos técnicos e processuais e respeitar as liberdades civis.

8. Treinamento de IA para a Lei de Aquisição de Força de Trabalho, outubro de 2022

O AI Training for the Acquisition Workforce Act determina o desenvolvimento e a implementação de um programa de treinamento de IA para pessoal designado no governo federal. O programa visa fornecer conhecimento abrangente sobre capacidades de IA, riscos associados e estratégias de mitigação, com componentes de aprendizagem interativos e atualizações regulares para garantir eficácia e alinhamento com os últimos desenvolvimentos de IA.

9. Projeto para uma Declaração de Direitos da IA publicado pelo Escritório de Política Científica e Tecnológica (OSTP), outubro de 2022

O Blueprint for an AI Bill of Rights é um conjunto de cinco princípios e práticas associadas para ajudar a orientar o design, uso e implantação de sistemas automatizados para proteger os direitos dos americanos. Esses cinco princípios são sistemas seguros e eficazes, proteções de discriminação algorítmica, privacidade de dados, notificação e explicação, e alternativas humanas, consideração e fallback.

10. NIST Secure Software Development Framework (SSDF V1.1): Recomendações para mitigar o risco de vulnerabilidades de software

O NIST SSDF descreve um conjunto básico de práticas de desenvolvimento de software seguro de alto nível que podem ser integradas em implementações de SDLC, o que preenche uma lacuna nos recursos tradicionais de SDLC que não abordam explicitamente a segurança do software.
Prepare the Organization (PO), Protect the Software (PS), Produce Well-Secured Software (PW) e Respond to Vulnerabilities (RV). Cada prática inclui quatro subelementos: Practice, Tasks, Notional Implementation Examples e References.

11. Estrutura de responsabilidade da IA para agências federais, publicado pelo GAO, junho de 2021

O AI Accountability Framework for Federal Agencies é composto por quatro princípios (governança, dados, desempenho e monitoramento). O framework captura práticas de accountability essenciais centradas nos quatro princípios para ajudar agências federais a usar a IA de forma responsável. Este framework fornece um conjunto abrangente de diretrizes e melhores práticas para garantir transparência, justiça e accountability no desenvolvimento e implantação de sistemas de IA, ao mesmo tempo em que considera preocupações com privacidade e segurança.

12. Lei da Iniciativa Nacional de IA de 2020, promulgada em janeiro de 2021

O National Artificial Intelligence Initiative Act de 2020 foi promulgado em janeiro de 2021 como parte do National Defense Authorization Act. Ele estabeleceu o National AI Initiative Office via OSTP. O National Science and Technology Council também foi encarregado de criar um comitê interinstitucional para coordenar programas e atividades federais para apoiar a iniciativa e um comitê consultivo para o presidente, e exige um estudo sobre o impacto da IA na força de trabalho dos EUA. O ato também exige pesquisa adicional de IA pelo GAO para que a NSF forneça subsídios para outras pesquisas de IA e padrões voluntários de IA pelo NIST.

13. Ordem Executiva 13960: Promovendo o uso de IA confiável no governo federal, dezembro de 2020

EO 13960 enfatiza o potencial da IA para aprimorar as operações governamentais e pede sua implementação responsável e confiável, definindo diretrizes para agências. A ordem lista nove características de IA confiável que devem ser consideradas ao longo do ciclo de vida da IA para garantir o design, desenvolvimento, implantação e operacionalização confiáveis da IA.

Nove princípios de IA confiável conforme enumerados na EO 13960

1. Legal e respeitoso dos valores da nossa nação 2. Proposital e orientado para o desempenho
3. Preciso, confiável e eficaz 4. Seguro, protegido e resiliente
5. Compreensível
6. Responsável e rastreável 7. Monitorado regularmente 8. Transparente
9. Responsável

14. Lei de IA no Governo, setembro de 2020 A Lei de IA no Governo de 2020 estabeleceu o Centro de Excelência em IA (AI CoE) dentro da Administração de Serviços Gerais para facilitar a adoção de IA e melhorar as operações e competências do governo.
A lei promove a colaboração entre agências, indústria, organizações sem fins lucrativos e instituições educacionais para promover a adoção de IA e inclui orientação do Diretor do Escritório de Gestão e Orçamento para proteger as liberdades civis e a segurança nacional no uso da tecnologia de IA.

15. Ordem Executiva 13859: Mantendo a Liderança Americana em IA, fevereiro de 2019 A ordem executiva enfatiza a necessidade do avanço da IA em todo o Governo federal, indústria e academia para aproveitar seu potencial para os americanos. A ordem reconhece a importância da GenAI e descreve uma estratégia abrangente para promover seu desenvolvimento, implantação e regulamentação, ao mesmo tempo em que salvaguarda o interesse e os valores nacionais.

16. Lei de Autorização de Defesa Nacional John S. McCain, Seção 1051 para o Ano Fiscal de 2019

A Lei de Autorização de Defesa Nacional (NDAA) define inteligência artificial para incluir cada um dos seguintes:

- ÿ Qualquer sistema artificial que executa tarefas em circunstâncias variadas e imprevisíveis sem supervisão humana significativa, ou que pode aprender com a experiência e melhorar o desempenho quando exposto a conjuntos de dados.
- ÿ Um sistema artificial desenvolvido em software de computador, hardware físico ou outro contexto que resolve tarefas que exigem percepção, cognição, planejamento, aprendizado, comunicação ou ação física semelhantes às humanas.
- ÿ Um sistema artificial projetado para pensar ou agir como um humano, incluindo arquiteturas cognitivas e redes neurais.
- ÿ Um conjunto de técnicas, incluindo aprendizado de máquina, projetadas para aproximar uma tarefa cognitiva.

ÿ Um sistema artificial projetado para agir racionalmente, incluindo um agente de software inteligente ou incorporado robô que atinge objetivos usando percepção, planejamento, raciocínio, aprendizado, comunicação, tomada de decisão e ação.

17. Lei do Governo Eletrônico de 2002

O E-Government Act de 2002 é um estatuto dos EUA que busca "melhorar a gestão e a promoção de serviços e processos de governo eletrônico, estabelecendo um Diretor de Informação federal dentro do Escritório de Gestão e Orçamento e estabelecendo uma estrutura de medidas que exigem o uso de tecnologia da informação baseada na Internet para melhorar o acesso dos cidadãos às informações e serviços do governo e para outros fins". 62 Ele também exige a criação de Avaliações de Impacto à Privacidade ("PIAs") para uso em todas as agências federais que criam novas tecnologias que gerenciam informações identificáveis.

Apêndice F. Relatórios sobre IA em relação direta com a pesquisa e desenvolvimento (P&D) para a ciência

A seguir estão referências selecionadas de relatórios, descobertas e identificação de necessidades em IA relacionadas à pesquisa e desenvolvimento, que orientam o uso de sistemas de IA por seus conjuntos exclusivos de princípios e considerações para o avanço da ciência.

1. "Relatório de IA para Ciência, Energia e Segurança (AI4SES)" Jonathan Carter, John Feddema, Doug Kothe, Rob Neely, Jason Pruet e Rick Stevens, 2023
2. "Plano Estratégico Nacional de Pesquisa e Desenvolvimento em Inteligência Artificial", Comissão Selecta sobre IA, Conselho Nacional de Ciência e Tecnologia, 2023 3. "Inteligência Artificial para Isótopos: Relatório sobre o Workshop de 2022 sobre Inteligência Artificial para P&D e Produção de Isótopos", Kristian Myhre, Draguna Vrabie, Danda Rawat e Ethan Balkin, 2023 4. "Traçando um caminho em um cenário técnico e geopolítico em mudança: computação pós-exaescala para a Administração Nacional de Segurança Nuclear ," Academias Nacionais de Ciências, Engenharia e Medicina, 2023

5. "Relatório Executivo Interino do AI@DOE," Ray Grout, Kelly Rose, Valerie Taylor e Brian Essen, 2022 6. "Inteligência Artificial para Previsibilidade do Sistema Terrestre (AI4ESP)" Nicki Hickmon, et al, 2022 7. "Gerenciamento e armazenamento de dados científicos (para fluxos de trabalho de IA/ML)," Suren Byna, et al, 2022 8. "Inteligência Artificial e Aprendizado de Máquina para Oportunidades e Desafios de Pesquisa em Bioenergia", Huimin Zhao, Nathan Hillson, Kerstin Kleese van Dam, Deepti Tanjore, e outros, 2022 ()
9. "Relatório de IA para a Ciência" Rick Stevens, Valerie Taylor, Jeff Nichols, Arthur Maccabe, Katherine Yelick e David Brown, 2020

10. "Rascunho final do relatório ao comitê pelo Subcomitê de IA/ML, ciência intensiva em dados e computação de alto desempenho", Comitê Consultivo de Computação Científica Avançada do DOE (ASCAC), 2020

11. "Oportunidades e desafios da inteligência artificial e da aprendizagem de máquina para o avanço de Ciência, Tecnologia e do Escritório de Missões Científicas", Tony Hey, et al, 2020 12. "Aplicação de Inteligência Artificial e Aprendizado de Máquina para a Operação de Instalações de Aceleradores de NP: Reunião da Mesa Redonda de NP," DOE Office of Science, Nuclear Physics, 2020 13. "IA de próxima geração para detecção de proliferação: acelerando o desenvolvimento e o uso da explicabilidade Métodos para projetar sistemas de IA adequados para aplicações de missão de não proliferação: Relatório de workshop," Francisco Alexander, e outros, 2021
14. "IA para Física Nuclear: Workshop do Escritório de Ciências do DOE", Tanja Horn, et al, 2020 15. "IA para Física Nuclear: Relatório do Workshop do Escritório de Ciências do DOE," Paulo Bedaque, et al, 2020 16. "Avançando na fusão com aprendizado de máquina: relatório do Office of Science", DOE Ciências de Energia de Fusão e Pesquisa em Computação Científica Avançada, 2019
17. "Produzindo e gerenciando grandes dados científicos com inteligência artificial e aprendizado de máquina: Relatório da Mesa Redonda do DOE," Daniel Ratner e Bobby Sumpter, 2019

18. “Dados e modelos - Uma estrutura para o avanço da IA na ciência: Relatório da mesa redonda do DOE”, Kjiersten Fagnan, et al, 2019
19. “Necessidades básicas de pesquisa para aprendizado de máquina científica: tecnologias essenciais para inteligência artificial: Relatório do Workshop do DOE,” Nathan Baker e outros, 2019
20. “Aprendizado e compreensão de máquina para computação científica inteligente em escala extrema e Descoberta: Relatório do Workshop do DOE,” Michael Berry, Thomas Potok, Prasanna Balaprakash, Hank Hoffman, Raju Vatsavai e Prabhat, 2015

Apêndice G. O ciclo de vida da IA

O ciclo de vida da inteligência artificial inclui três fases: **pré-design, design e desenvolvimento** e **Implantação.** Veja a Figura 7 para uma descrição das etapas contidas em cada fase e seus riscos associados.

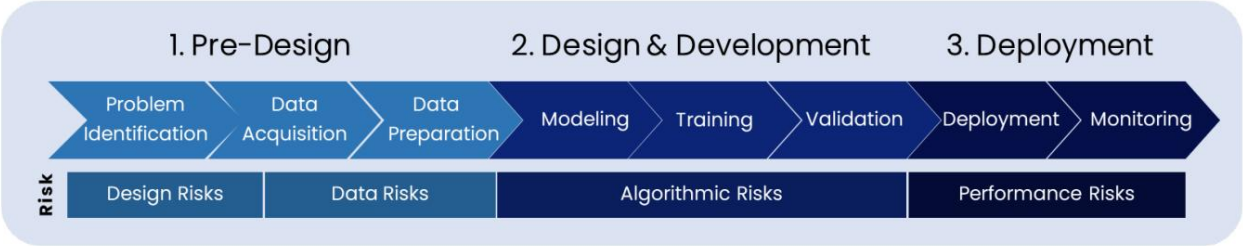


Figura 7: O Ciclo de Vida da IA e Riscos Associados. Adaptado de: The Department of Energy Artificial Intelligence Resource

Cada uma das subfases contém ações críticas necessárias para atingir o resultado desejado ao preparar, desenvolver e implantar uma nova tecnologia de IA. Publicações recentes, como a EO 14110, colocam foco adicional na Fase Dois: Design e Desenvolvimento.63 Incluídas nesta fase estão atividades relacionadas ao teste de um modelo de IA. O teste de IA envolve o uso de ferramentas, estruturas e metodologias apropriadas para validar independentemente os modelos e sua integração com sistemas e processos upstream e downstream existentes. O objetivo de concluir testes de IA robustos é garantir que os padrões de qualidade, segurança e conformidade sejam atendidos e que a solução seja escalável e confiável. O teste de IA é um ciclo iterativo em si mesmo e requer monitoramento, feedback e melhoria consistentes. As atividades de teste devem cobrir defeitos na funcionalidade, interações do usuário com a ferramenta e a experiência do usuário, entre outros requisitos específicos da ferramenta. Além disso, o teste pode ajudar a identificar dentro de um modelo os riscos específicos do GenAI discutidos na Seção 8.

De acordo com a EO 14110, orientações específicas sobre o desenvolvimento de tecnologias GenAI estão sendo escritas para complementar a Estrutura de Desenvolvimento de Software Seguro (SSDF) do NIST existente. Além disso, espera-se que o red-teaming em colaboração com o NIST, que envolve “um grupo de pessoas autorizadas e organizadas para emular os recursos de ataque ou exploração de um adversário em potencial contra a postura de segurança de uma empresa”, desempenhe um papel fundamental no teste e monitoramento dos recursos do GenAI durante as Fases Dois e Três, conforme discutido na EO 14110 Seção 10.1 (b)(vii)(A). [64] Incluindo red-teaming no desenvolvimento de IA O ciclo de vida pode melhorar a segurança do modelo demonstrando tanto os impactos de ataques bem-sucedidos quanto possíveis ações responsivas ou remediações que sejam eficazes contra tais ataques.

Cada fase possui um conjunto único de riscos que devem ser abordados ou mitigados durante essa parte do ciclo de vida. Para obter informações adicionais sobre a mitigação de riscos ao longo do ciclo de vida da IA, consulte a Seção 8.3: Gerenciamento de Riscos de IA e o Manual de Gerenciamento de Riscos de IA (AIRMP). O AIRMP é um conjunto de 141 riscos relacionados ao desenvolvimento e à implantação de IA, juntamente com mitigações correspondentes que podem ser incorporadas proativamente ao longo do ciclo de vida da IA para abordar riscos de design, dados, algoritmos e desempenho.

Apêndice H. Estrutura de Gestão de Riscos de Inteligência Artificial do NIST (NIST AI RMF Versão 1.0)

Além de fornecer informações valiosas sobre os princípios de IA confiável e responsável e como o gerenciamento de riscos eficaz pode viabilizar esses princípios, o NIST AI RMF inclui uma estrutura voluntária para gerenciar riscos de IA, composta por quatro funções: Governar, Mapear, Medir e Gerenciar.

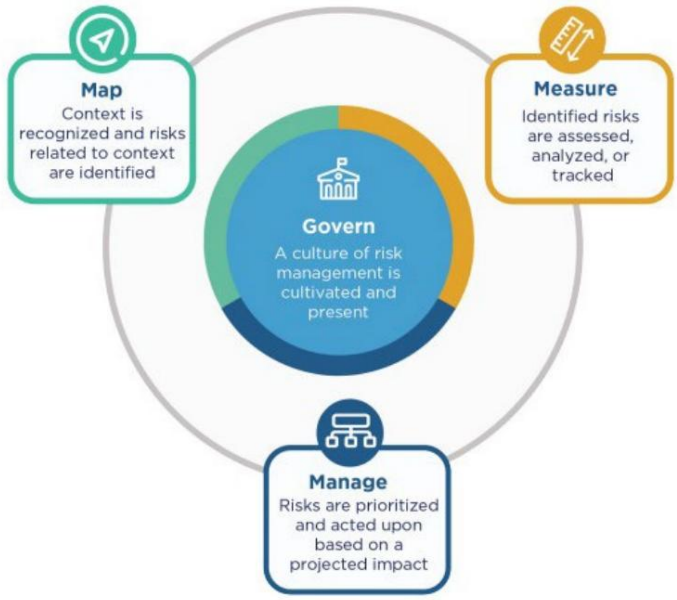


Figura 8: As quatro funções do NIST AI RMF

A função **Govern** facilita o restante das funções cultivando e institucionalizando uma cultura de gerenciamento de risco em toda a organização. Isso inclui desenvolver uma estrutura de governança e processos e documentação correspondentes para avaliar, identificar e gerenciar riscos de IA. A função Govern também conecta o desenvolvimento do sistema de IA a princípios organizacionais, apetites de risco (a tolerância ou disposição da organização em aceitar riscos) e estratégias.

A função **Mapa** cria o contexto adequado para enquadrar riscos relacionados a um sistema de IA. As informações obtidas por meio da conclusão da função Map permitem a seleção e a implantação de casos de uso de IA apropriados e previnem ou mitigam riscos. A obtenção do contexto organizacional permite a compreensão mais completa dos riscos de IA e fatores contribuintes ou subjacentes. Os resultados da função Map também facilitam as funções Measure e Manage.

A função **Measure** utiliza uma variedade de ferramentas e métodos para “analisar, avaliar, comparar e monitorar o risco de IA e impactos relacionados”, incluindo a documentação da funcionalidade e da confiabilidade potencial de um sistema. A função Measure depende de dados da função Map e informa a função Manage.

A função **Manage** facilita a alocação de recursos para riscos conforme definido pela função Govern por meio de um plano de tratamento de risco. As ações podem incluir resposta, recuperação e comunicação no caso de um incidente técnico ou de segurança. Informações de impactos negativos podem ajudar a informar futuras decisões de recursos.

Para obter informações adicionais sobre ações específicas para cada função, consulte a publicação online completa do NIST AI RMF, ou para obter informações sobre gerenciamento geral de risco de IA, consulte a Seção 8.3.

Apêndice I. Engenharia de Prompt GenAI

Entender como elaborar corretamente prompts (o termo técnico para uma solicitação, comando ou pergunta) ao interagir com um modelo GenAI é fundamental para obter uma saída confiável e precisa.

O primeiro princípio da elaboração de prompts de IA é escrever instruções claras e específicas. Escrever um prompt claro e um prompt curto não são sinônimos, geralmente prompts mais longos são mais específicos e levam a melhores resultados. Usar delimitadores como aspas ou acentos graves para esclarecer para o modelo qual texto tomar como entrada (para resumir, expandir ou transformar de alguma forma) versus qual texto está transmitindo as ações

para o modelo executar (resumir, expandir, etc.) pode garantir uma melhor resposta e evitar injeções imediatas que poderiam alterar o propósito do modelo.

Outra tática para fornecer instruções claras e específicas é solicitar uma resposta padronizada da IA no prompt (por exemplo, solicitando uma resposta HTML ou JSON para a consulta). Por fim, pode ser útil incluir no prompt uma solicitação para que a IA verifique se as condições para uma resposta foram satisfeitas ou, se não forem, para verificar suposições em vez de executar uma consulta completa. Por exemplo, se estiver alimentando o modelo com texto do qual extrair instruções passo a passo, dê ao modelo uma segunda opção de resposta se nenhum conteúdo para criar etapas for detectado.

O segundo princípio para elaborar prompts de IA é dar tempo ao modelo para pensar, ou dedicar mais poder computacional para resolver um problema para evitar erros. Uma tática para implementar esse princípio é pedir ao modelo para executar a consulta em etapas para garantir que cada parte da tarefa seja concluída completamente. Essa tática se alinha bem com a tática de fornecer uma saída padronizada. Outra tática para dar tempo ao modelo para pensar é pedir que ele calcule sua própria solução para um problema antes de verificar a correção de uma solução fornecida.

Os prompts são desenvolvidos de forma mais eficaz ao incorporar esses princípios em um ciclo de desenvolvimento iterativo. O primeiro passo no ciclo é ter a ideia do que você gostaria de pedir para o modelo fazer. Em seguida, vem uma implementação inicial, ou uma primeira passagem de configuração do ambiente e escrita de uma consulta.
Executar essa consulta fornecerá um resultado experimental que será usado para analisar quais erros o modelo está cometendo ou quais informações fornecer no prompt para aprimorar melhor a saída para seu propósito pretendido. Essa análise de erro é então usada para criar a ideia para o prompt atualizado, e o ciclo se repete.

Este conteúdo é um resumo de um curso OpenAI sobre engenharia de prompts e funcionalidades de IA generativa.65 O curso completo e os materiais relacionados podem ser encontrados aqui. Informações adicionais da OpenAI sobre engenharia rápida podem ser encontradas aqui.

Apêndice J. Exemplos de dados protegidos

Tipos de dados inerentemente sensíveis	Descrição
Informação classificada	Informações classificadas são definidas pelo DOE como “certas informações que requerem proteção contra divulgação não autorizada no interesse da defesa e segurança nacional ou relações exteriores dos Estados Unidos, de acordo com estatuto federal ou ordem executiva”. Inclui Dados Restritos, Dados Anteriormente Restritos e Informações de Segurança Nacional. O dano potencial à segurança nacional de cada um é denotado pelos níveis de classificação Top Secret, Secret e Confidential. **Fonte:** Informação Classificada (DOE)
Informações proprietárias da empresa	O NIST define informações proprietárias como “materiais e informações relacionadas ou associadas aos produtos, negócios ou atividades de uma empresa, incluindo, mas não se limitando a, informações financeiras, dados ou declarações, segredos comerciais, pesquisa e desenvolvimento de produtos, designs de produtos existentes e futuros e especificações de desempenho, planos ou técnicas de marketing, esquemas, listas de clientes, programas de computador, processos e know-how que foram claramente identificados e devidamente marcados pela empresa como informações proprietárias, segredos comerciais ou informações confidenciais da empresa. As informações devem ter sido desenvolvidas pela empresa e não estar disponíveis ao Governo ou ao público sem restrição de outra fonte.” **Fonte:** Definição de Informação Proprietária (NIST)
Controlado não classificado informação (CUI)	CUI é definido pelo DOE como “informação que o Governo cria ou possui, ou que uma entidade cria ou possui para ou em nome do Governo”. Governo, que uma Lei, Regulamento ou Política Governamental (LRGWP)

Tipos de dados inerentemente sensíveis	Descrição
	exige ou permite que uma agência lide com o uso de controles de proteção ou disseminação." Isso inclui informações confidenciais, informações de identificação pessoal (PII), informações confidenciais e informações privadas. **Fonte:** Informações não classificadas controladas (doe.gov)
Informação confidencial	*Para fins deste Guia,* informações confidenciais podem ser definidas de forma geral como informações ou dados de natureza pessoal que são propriedade de um indivíduo, ou informações ou dados pertencentes ou enviados por uma organização. No entanto, informações confidenciais referem-se a quaisquer dados ou conhecimentos que são compartilhados com um indivíduo ou organização sob a condição de que permaneçam privado e não divulgado. Inclui todas as informações que são designadas como confidenciais ou que, por sua natureza, são ou devem ser consideradas confidenciais (se estiver assim marcado) e inclui todos os dados pessoais (PII) e informações proprietárias que se relacionam com a propriedade intelectual, dados, know-how, segredos comerciais, negócios, desenvolvimentos, pessoal e fornecedores da entidade à qual as informações pertencem. As informações confidenciais existem em todas as formas: escritas, faladas, observadas, eletrônicas ou de outra forma. Se houver qualquer incerteza ou questão legal relacionada à confidencialidade das informações ou dados, deve-se procurar aconselhamento do DOE ou do consultor jurídico do contratado antes do uso. **Exemplo de fonte:** 452.224-70 Confidencialidade das informações. | Acquisition.GOV
Pesquisa Cooperativa e Acordo de Desenvolvimento (CRADA)-informações protegidas	O DOE define informações protegidas do CRADA como "informações geradas que são marcadas como Informações Protegidas do CRADA por uma Parte deste CRADA e que seriam Informações Proprietárias se tivessem sido obtidas de uma entidade não federal". Esse tipo de informação é protegida por cinco anos por lei e, portanto, é protegida por um período de cinco a 30 anos. **Fonte:** DOE O 483.1B Acordos de Pesquisa e Desenvolvimento Cooperativo
Propriedade intelectual (PI)	Propriedade Intelectual inclui segredos comerciais, patentes, direitos autorais, marcas registradas, desenhos industriais e indicações geográficas. **Fonte:** Organização Mundial da Propriedade Intelectual
Dados de direitos limitados	Dados nos quais o governo dos EUA não tem direitos inerentes. Dados de direitos limitados são normalmente desenvolvidos com despesas privadas, não sob uma concessão, e são normalmente utilizáveis apenas para os propósitos de uma concessão específica. Dados de direitos limitados devem ser devidamente marcados como destinatário e sua entrega deve ser minimizada. Para usar ou compartilhar dados de direitos limitados fora do governo dos EUA, permissão por escrito deve ser obtida do proprietário dos dados.
Informações de identificação pessoal (PII)	Informações de identificação pessoal (PII) são definidas pela Circular nº A–130 do Escritório de Administração e Orçamento (OMB) (e definidas na Ordem Executiva 14110 usando esta fonte) como "informações que podem ser usadas para distinguir ou rastrear a identidade de um indivíduo, sozinhas ou quando combinadas com outras informações vinculadas ou vinculáveis a um indivíduo específico". **Fonte:** Circular nº A–130 do Gabinete de Gestão e Orçamento (OMB) As informações de identificação pessoal (PII) são definidas pelo DOE como: "Informações que podem ser usadas para distinguir ou rastrear a identidade de um indivíduo, seja sozinha ou quando combinadas com outras informações que são vinculadas ou vinculáveis a um indivíduo específico. PII pode incluir identificadores individuais exclusivos ou combinações de identificadores, como o nome de um indivíduo, número do Seguro Social, data e local de nascimento, nome de solteira da mãe, dados biométricos, etc.

Tipos de dados inerentemente sensíveis	Descrição
	A sensibilidade do PII aumenta quando combinações de elementos aumentam a capacidade de identificar ou atingir um indivíduo específico. O PII, que se perdido, comprometido ou divulgado sem autorização, pode resultar em dano substancial, constrangimento, inconveniência ou injustiça para um indivíduo, é categorizado como PII de Alto Risco. Exemplos de PII de Alto Risco incluem Números de Previdência Social (SSNs), registros biométricos (por exemplo, impressões digitais, DNA, etc.), informações médicas e de saúde, informações financeiras (por exemplo, números de cartão de crédito, relatórios de crédito, números de contas bancárias, etc.) e informações de segurança (por exemplo, informações de autorização de segurança). Embora todas as PII devam ser tratadas e protegidas adequadamente, as PII de alto risco devem receber maior proteção e consideração após uma violação, devido ao risco aumentado de danos a um indivíduo se forem mal utilizadas ou comprometidas." **Fonte:** DOE O 206.1, Programa de Privacidade do Departamento de Energia, 1º de janeiro de 2009
Informações privadas	A maioria das definições de "informações privadas" existe no nível estadual, local ou organizacional e contém similaridades com as definições federais de informações confidenciais de PII, significando em geral, qualquer informação referente a uma pessoa física que, por causa de um identificador, pode ser usada para identificar tal pessoa física se estiver em combinação com qualquer um ou mais elementos de dados específicos, como identificados na definição de PII. Se "informações privadas" for um tipo específico de informação dentro de sua organização ou escritório, consulte essa definição e quaisquer requisitos associados em torno de sua confidencialidade, integridade e disponibilidade. Se houver alguma incerteza ou questão legal sobre se a informação é informação privada, deve-se procurar aconselhamento do DOE ou do advogado contratado antes do uso.
Dados de aquisição (dados de preço ou custo)	*Dados de custo ou preço* (10 USC 3701(1) e 41 USC capítulo 35) significa todos os fatos que, na data do acordo de preço, ou, se aplicável, uma data anterior acordada entre as partes que seja o mais próximo possível da data do acordo sobre o preço, compradores e vendedores prudentes esperariam razoavelmente que afetassem as negociações de preço significativamente. Os dados de custo ou preço são factuais, não de julgamento; e são verificáveis. Embora não indiquem a precisão do julgamento do contratante em potencial sobre os custos ou projeções futuras estimadas, eles incluem os dados que formam a base para esse julgamento. Os dados de custo ou preço são mais do que dados contábeis históricos; são todos os fatos que podem ser razoavelmente esperados para contribuir para a solidez das estimativas de custos futuros e para a validade das determinações de custos já incorridos. Eles também incluem, mas não estão limitados a, fatores como- (1) Cotações de fornecedores; (2) Custos não recorrentes; (3) Informações sobre alterações nos métodos de produção e no volume de produção ou de compras; (4) Dados que suportem as projeções das perspectivas e objetivos do negócio e dos custos operacionais relacionados; (5) Tendências de custos unitários, como as associadas à eficiência do trabalho; (6) Decisões de fazer ou comprar; (7) Recursos estimados para atingir os objetivos do negócio; e (8) Informações sobre decisões de gestão que podem ter um impacto significativo nos custos. *Dados que não sejam dados de custo ou de preços certificados* significam dados de preços, dados de custo e informações de julgamento necessárias para que o responsável pela contratação determine

Tipos de dados inerentemente sensíveis	Descrição
	um preço justo e razoável ou para determinar o realismo de custos. Tais dados podem incluir os mesmos tipos de dados que os dados de custo ou preço certificados, consistentes com a Tabela 15-2 de 15.408, mas sem a certificação. Os dados também podem incluir, por exemplo, dados de vendas e quaisquer informações razoavelmente necessárias para explicar o processo de estimativa do proponente, incluindo, mas não se limitando a - (1) Os fatores de julgamento aplicados e os métodos matemáticos ou outros usados na estimativa, incluindo aqueles usados na projeção a partir de dados conhecidos; e (2) A natureza e o montante de quaisquer contingências incluídas no preço proposto. **Fonte:** Parte 2 - Definições de palavras e termos | Acquisition.GOV
Informações de saúde protegidas (PHI)	PHI é definido pela Regra de Privacidade da Lei de Portabilidade e Responsabilidade de Seguros de Saúde (HIPAA) como “informações de saúde individualmente identificáveis que são transmitidas ou mantidas em qualquer formato ou meio (eletrônico, oral ou papel) por uma entidade coberta ou seus associados comerciais, excluindo certos registros educacionais e de emprego”. **Fonte:** Privacy Rule and Research (NIH)
Informação sensível	O Manual de Segurança de Operações (OPSEC) do DOE 2019, Seção 3.1.5 “Revisão de Divulgação Pública” discute duas características primárias de uma informação que determinam se essa informação é segura para divulgação pública. Essas duas características são: sensibilidade e risco. Juntos, esses dois termos sugerem que o termo “informação sensível” é um termo categórico, que então inclui outros tipos específicos de informação sensível. O Manual se aproxima de uma definição para informações sensíveis na Seção 3.1.5 ao apresentar estas características primárias para “determinar a adequação para divulgação” de informações: ÿ Sensibilidade: “Se a informação for divulgada ao público, ela não deve revelar ou identificar informações, atividades ou programas sensíveis.” ÿ Risco: “Informações que podem ser usadas por adversários em detrimento de funcionários, do público, do departamento ou da nação não devem ser aprovadas para divulgação. Essa determinação deve ser baseada em princípios sólidos de gerenciamento de risco focados na prevenção de potenciais consequências adversas.” **Fonte:** Manual de Operações e Segurança do DOE (OPSEC), 2019

Apêndice K. Glossário

Há uma variedade de termos relevantes para a discussão em torno de IA e GenAI. As definições listadas abaixo são uma seleção de termos que podem ser úteis ao realizar pesquisa, design, discussão, operação ou desenvolvimento. Todas as definições são originadas da Seção 3 da Ordem Executiva 14110 sobre *Desenvolvimento e uso seguro, protegido e confiável de inteligência artificial,* a menos que marcado * e citado.

Prazo	Definição
Inteligência artificial (IA)	Um sistema baseado em máquina que pode, para um dado conjunto de objetivos definidos por humanos, fazer previsões, recomendações ou decisões que influenciam ambientes reais ou virtuais. Sistemas de inteligência artificial usam entradas baseadas em máquinas e humanos para perceber ambientes reais e virtuais; abstrair tais percepções em modelos por meio de análise de forma automatizada; e usar inferência de modelo para formular opções de informação ou ação.

Prazo	Definição
Inteligência artificial generativa (IA generativa ou (GenAI)	A classe de modelos de IA que emulam a estrutura e as características de dados de entrada para gerar conteúdo sintético derivado. Isso pode incluir imagens, vídeos, áudio, texto e outros conteúdos digitais.
Modelo de IA	Um componente de um sistema de informação que implementa tecnologia de IA e usa técnicas computacionais, estatísticas ou de aprendizado de máquina para produzir saídas a partir de um determinado conjunto de entradas.
Red-teaming de IA	Um esforço de teste estruturado para encontrar falhas e vulnerabilidades em um sistema de IA, geralmente em um ambiente controlado e em colaboração com desenvolvedores de IA. A formação de equipes vermelhas de inteligência artificial é mais frequentemente realizada por “equipes vermelhas” dedicadas que adotam métodos adversários para identificar falhas e vulnerabilidades, como resultados prejudiciais ou discriminatórios de um sistema de IA, comportamentos imprevistos ou indesejáveis do sistema, limitações ou riscos potenciais associados ao uso indevido do sistema.
Sistema de IA	Qualquer sistema de dados, software, hardware, aplicativo, ferramenta ou utilitário que opere total ou parcialmente usando IA.
Tecnologias críticas e emergentes	As tecnologias listadas na atualização da lista de tecnologias críticas e emergentes de fevereiro de 2022 emitido pelo Conselho Nacional de Ciência e Tecnologia (NSTC), conforme alterado por atualizações subsequentes da lista emitida pelo NSTC.
Aprendizado profundo (DL)*	Deep learning é um subconjunto do machine learning, e que é essencialmente uma rede neural com três ou mais camadas. Essas redes neurais tentam simular o comportamento do cérebro humano — embora longe de corresponder à sua capacidade — permitindo que eles “aprendam” com grandes quantidades de dados. Enquanto uma rede neural com uma única camada ainda pode fazer previsões aproximadas, camadas ocultas adicionais podem ajudar a otimizar e refinar para precisão.66
Rede neural profunda (DNN)*	Redes neurais profundas consistem em múltiplas camadas de nós interconectados, cada um construindo sobre a camada anterior para refinar e otimizar a predição ou categorização. Essa progressão de computações através da rede é chamada de propagação para frente. As camadas de entrada e saída de uma rede neural profunda são chamadas de camadas visíveis. A camada de entrada é onde o modelo de aprendizado profundo ingere os dados para processamento, e a camada de saída é onde a predição ou classificação final é feita.67
Modelo de fundação de dupla utilização	Um modelo de IA treinado em dados amplos; geralmente usa autossupervisão; contém pelo menos dezenas de bilhões de parâmetros; é aplicável em uma ampla gama de contextos; e que exibe, ou pode ser facilmente modificado para exibir, altos níveis de desempenho em tarefas que representam um sério risco à segurança, à segurança econômica nacional, à saúde ou segurança pública nacional, ou qualquer combinação desses assuntos, como por exemplo: ÿ reduzir substancialmente a barreira de entrada para não especialistas no design, sintetizar, adquirir ou utilizar produtos químicos, biológicos, radiológicos ou nucleares (armas QBRN) ÿ possibilitar operações cibernéticas ofensivas poderosas por meio de operações automatizadas descoberta e exploração de vulnerabilidades contra uma ampla gama de alvos potenciais de ataques cibernéticos ÿ permitir a evasão do controlo ou supervisão humana através de meios engano ou ofuscação

Prazo	Definição
	Os modelos atendem a essa definição mesmo que sejam fornecidos aos usuários finais com salvaguardas técnicas que tentam impedir que os usuários tirem proveito dos recursos inseguros relevantes.
GPT*	GPT é uma família de LLMs construída na arquitetura de rede neural profunda (DNN) que foi ajustada usando técnicas de processamento de linguagem natural (NLP) e aprendizado por reforço de feedback humano (RLHF).
Modelo de linguagem grande (LLM)*	Grandes modelos de linguagem (LLMs) aproveitam o aprendizado autossupervisionado e podem aprender com grandes quantidades de dados de texto não estruturados e não rotulados. Esses modelos são treinados em grandes conjuntos de dados, permitindo que um modelo seja usado para vários casos de uso.
Aprendizado de máquina	Um conjunto de técnicas que podem ser usadas para treinar algoritmos de IA para melhorar o desempenho em uma tarefa com base em dados.
Processamento de linguagem natural (PNL)*	O processamento de linguagem natural (PLN) refere-se ao ramo da ciência da computação: e, mais especificamente, o ramo da inteligência artificial ou IA — preocupado em dar aos computadores a capacidade de "entender" textos e palavras faladas da mesma forma que os seres humanos. A PNL combina linguística computacional — modelagem baseada em regras da linguagem humana — com modelos estatísticos, de aprendizado de máquina e de aprendizado profundo. Juntas, essas tecnologias permitem que os computadores processem a linguagem humana na forma de texto ou dados de voz e “compreendam” seu significado completo, incluindo a intenção e o sentimento do falante ou escritor.68
Aprendizagem por reforço com feedback humano (RLHF)*	Reinforcement learning from human feedback (RLHF) é uma técnica de machine learning (ML) que usa feedback humano para otimizar modelos de ML para autoaprendizagem mais eficiente. Técnicas de Reinforcement learning (RL) treinam software para tomar decisões que maximizem recompensas, tornando seus resultados mais precisos. O RLHF incorpora o feedback humano na função de recompensas, para que o modelo ML possa executar tarefas mais alinhadas com os objetivos, desejos e necessidades humanas.69
Conteúdo sintético	Informações, como imagens, vídeos, clipes de áudio e texto, que foram significativamente modificadas ou geradas por algoritmos, inclusive por IA.
Modelo específico de tarefa*	Toda a IA em vigor hoje é específica para tarefas, ou IA restrita. Esta é uma distinção importante, pois a capacidade geral de raciocinar, pensar e perceber é conhecida como Inteligência Artificial Geral (AGI), que, neste ponto, não é tecnicamente possível.70
Banco de testes	Uma instalação ou mecanismo equipado para conduzir testes rigorosos, transparentes e replicáveis de ferramentas e tecnologias, incluindo IA e PETs, para ajudar a avaliar a funcionalidade, usabilidade e desempenho dessas ferramentas ou tecnologias.
Transformador*	Transformadores de aprendizado profundo são um tipo de IA que são usados para aprender representações de dados de forma automatizada. Transformadores são projetados para lidar com dados sequenciais, como linguagem natural, tornando-os adequados para tarefas como classificação de texto, tradução automática e resposta a perguntas.71
Marca d'água	O ato de incorporar informações, que normalmente são difíceis de remover, em saídas criadas por IA — incluindo saídas como fotos, vídeos, clipes de áudio ou texto — com a finalidade de verificar a autenticidade da saída ou a identidade ou características de sua procedência, modificações ou transporte.

1 Guia de referência de IA generativa do DOE v1, publicado em setembro de 2023
2 "DOE Prepping Version 2 of GenAI Responsible Use Guide", MeriTalk, 7 de dezembro de 2023
3 Ordem Executiva (EO 14110) sobre o desenvolvimento e uso seguro, protegido e confiável de inteligência artificial Inteligência, 30 de outubro de 2023, Seção 3 (b)
4 Ordem Executiva (EO 14110) sobre o desenvolvimento e uso seguro, protegido e confiável de inteligência artificial Inteligência, 30 de outubro de 2023, Seção 3 (t)
5 Ordem Executiva (EO 14110) sobre o desenvolvimento e uso seguro, protegido e confiável de dispositivos artificiais Inteligência, 30 de outubro de 2023, Seção 3 (p)
6 "O que é aprendizado profundo?" IBM
7 Gartner, "Glossário de termos para IA generativa e modelos de grandes linguagens", Anthony Mullen, julho de 2023. GARTNER é uma marca registrada e marca de serviço da Gartner, Inc. e/ou suas afiliadas nos EUA e internacionalmente e é usada aqui com permissão. Todos os direitos reservados.
8 "NukeLM: Modelos de linguagem pré-treinados e ajustados para os domínios nuclear e energético", Universidade Cornell, 25 de maio de 2021
9 OpenAI apoiado pela Microsoft registra marca registrada para ChatGPT com tecnologia GPT-5
10 Gartner, "Prevê 2024: O futuro das tecnologias de IA generativas", Arun Chandrasekaran, Anthony Mullen, Lizzy Foo Kune, Nicole Greene, Jim Hare, Leinar Ramos, Anushree Verma, 28 de fevereiro de 2024
11 Ordem Executiva (EO 14110) sobre o desenvolvimento e uso seguro, protegido e confiável de dispositivos artificiais Inteligência, 30 de outubro de 2023, Seção 10.1(e)
12 "Aviso à comunidade de pesquisa: Uso de tecnologia de inteligência artificial generativa no processo de revisão de mérito da NSF", NSF - National Science Foundation, 14 de dezembro de 2023 13 Processo de Avaliação e Autorização de Segurança Empresarial da OCIO, agosto de 2023 14 NIST AI RMF 1.0 pág. 6 15 NIST AI RMF 1.0 pág. 8 16 NIST AI RMF 1.0 pg. 1 17 ISO/IEC TS 5723:2022, citado no NIST AI RMF 1.0 18 DOE EO 13960 Plano de Consistência 19 "Como fazer um Red Team de um modelo de IA de geração", Harvard Business Review, Andrew Burt, 4 de janeiro de 2024 20 CFR Título 10, Cap. X, Parte 1008 Registros do DOE mantidos sobre indivíduos (Lei de Privacidade)
21 "Princípios de Privacidade de Dados e Proteção de Informações para a Cidade de Portland", Cidade de Portland, adotado em 19 de junho de 2019
22 "Princípios de Privacidade de Dados e Proteção de Informações para a Cidade de Portland", Cidade de Portland, adotado em 19 de junho de 2019. 2019
23 "TechBriefing: ChatGPT, LLMs e IA generativa", Gartner, publicado em 20 de junho de 2023
24 Circular OMB nº A–130 Gestão de informações como um recurso estratégico, Escritório de Gestão e Orçamento, Julho de 2016
25 Ordem Executiva (EO 14110) sobre o desenvolvimento e uso seguro, protegido e confiável de dispositivos artificiais Inteligência, 30 de outubro de 2023, Seção 3 (j)
26 Ordem Executiva (EO 14110) sobre o desenvolvimento e uso seguro, protegido e confiável de dispositivos artificiais Inteligência, 30 de outubro de 2023, Seção 3 (z)
27 NIST AI RMF 1.0 pág. 17
28 Confidencialidade, glossário do NIST
29 Manual de Operações e Segurança do DOE (OPSEC), Seção 3.1.5, junho de 2019
30 Thaler v. Vidal Decisão do Tribunal Federal Veja também: Orientação sobre IA do Escritório de Patentes e Marcas dos EUA (USPTO)
31 Federal Register: Inteligência Artificial e Direitos Autorais
32 Federal Register / Vol. 89, No. 30 / Terça-feira, 13 de fevereiro de 2024 / Avisos
33 Federal Register :: Inteligência Artificial e Direitos Autorais 34 ISO/IEC TS 5723:2022 referenciado em NIST AI RMF 1.0 35 NIST AI RMF 1.0 pág. 14 36 NIST AI RMF 1.0 pág. 15 37 NIST AI RMF 1.0 pg. 17 38 Justiça e preconceito na inteligência artificial: uma breve pesquisa de fontes, impactos e estratégias de mitigação, Emilio Ferrara, Cornell University, 16 de abril de 2023 39 NIST AI RMF 1.0 pág. 18

40 Justiça e preconceito na inteligência artificial: uma breve pesquisa de fontes, impactos e estratégias de mitigação, Emilio Ferrara, Cornell University, 16 de abril de 2023 41 ISO 9000:2015, ISO 9000:2015(en), Sistemas de gestão da qualidade — Fundamentos e vocabulário 42 ISO/IEC TS 5723:2022 43 NIST AI RMF 1.0 pg. 16 44 Uma abordagem fundamentada: superando alucinações de IA", John Bohannon, 27 de julho de 2023 45 Uma abordagem fundamentada: superando alucinações de IA", John Bohannon, 27 de julho de 2023 46 Ordem Executiva (EO 14110) sobre o desenvolvimento e uso seguro, protegido e confiável de dispositivos artificiais Inteligência, 30 de outubro de 2023, Seção 10.1 (f)(iii)

47 Ordem Executiva (EO 14110) sobre o desenvolvimento e uso seguro, protegido e confiável de inteligência artificial Inteligência, 30 de outubro de 2023, Seção 4.1 (b)

48 DOE EO 13960 Plano de Consistência

49 Gartner, "Como pilotar IA generativa", Leinar Ramos, Anthony Mullen, Rajesh Kandaswamy, Radu Miclaus, Erick Brethenoux, Avivah Litan, Haritha Khandabattu, 10 de julho de 2023

50 DOE EO 13960 Plano de Consistência

51 DOE EO 13960 Plano de Consistência

52 DOE EO 13960 Plano de Consistência

53 DOE EO 13960 Plano de Consistência

54 "Aprendizado de máquina adversário: uma taxonomia e terminologia de ataques e mitigações", NIST, janeiro de 2024

55 DOE EO 13960 Plano de Consistência

56 "Uma abordagem sandbox para regular aplicações de inteligência artificial de alto risco", Universidade de Cambridge, 12 de novembro de 2021

57 O que são dados sintéticos e como eles podem ajudar você competitivamente, MIT Sloan School of Management, 23 de janeiro de 2019. 2023

58 "Métricas, Logs e Rastros: O Triângulo Dourado da Observabilidade no Monitoramento", DevOps, 8 de novembro de 2018

59 "Como fazer um Red Team de um modelo de IA de geração", Harvard Business Review, Andrew Burt, 4 de janeiro de 2024 60 S.1353 - 117º Congresso (2021-2022): Advancing American AI Act, Biblioteca do Congresso, dezembro de 2023 61 NIST SP 800-218, Secure Software Development Framework (SSDF) Versão 1.1: Recomendações para mitigar o risco de vulnerabilidades de software, fevereiro de 2022 62E-Gov Act de 2002, Digital.gov, dezembro de 2002 63 Ordem Executiva (EO 14110) sobre o desenvolvimento e uso seguro, protegido e confiável de inteligência artificial Inteligência, 30 de outubro de 2023, Seção 2 (a)

64 Red Team - Glossário, Instituto Nacional de Padrões e Tecnologia: CRSC 65 ChatGPT Prompt Engineering para desenvolvedores - DeepLearning.AI 66 O que é Deep Learning? | IBM

67 O que é Deep Learning? | IBM 68 O que é Processamento de Linguagem Natural? | IBM 69 O que é RLHF? - Aprendizado por reforço a partir do feedback humano explicado, Amazon Web Services

70 Terminologia-chave de IA | GSA - Centros de Excelência em Modernização de TI

71 O que são transformadores de aprendizagem profunda? - Google LaMDA