AI Technology Research

IGF Internet Governance Forum RIYADH 2024

Diretrizes para Desenvolvimento e Implantação Seguros de Sistemas de IA

kaspersky bring on the future

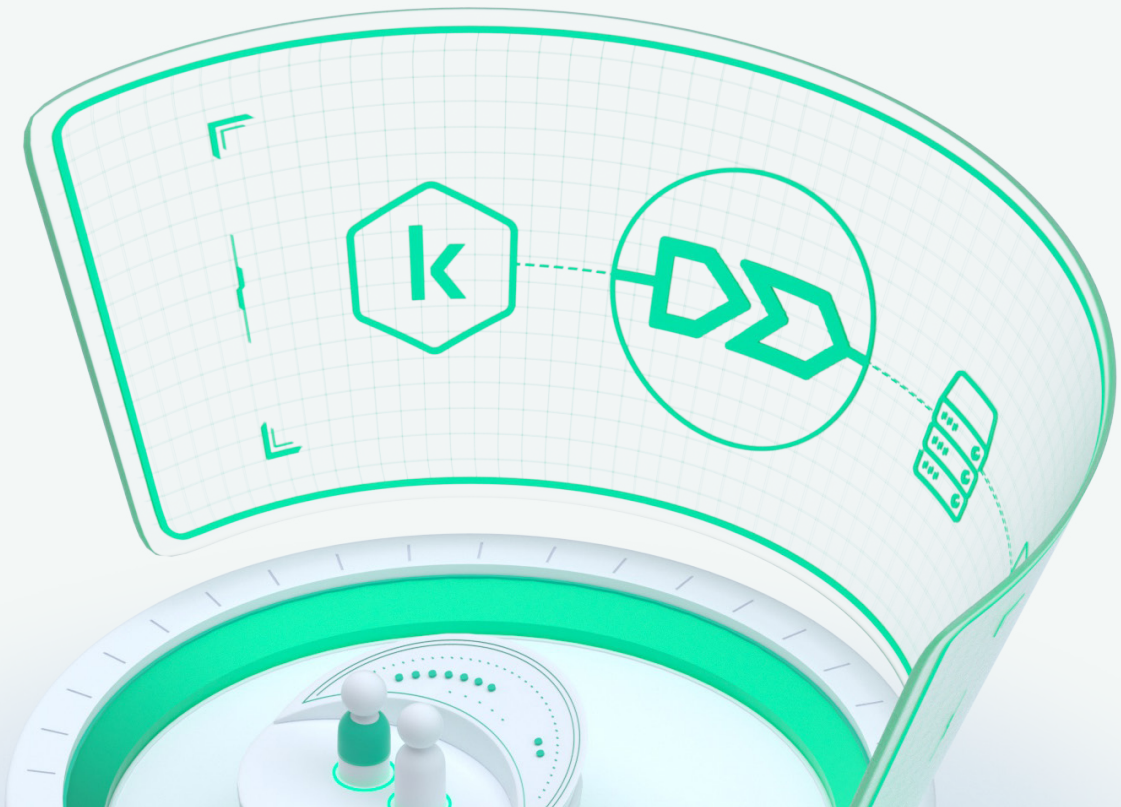
Agradecimentos

Este artigo foi desenvolvido pela Kaspersky e apresentado durante o **Workshop nº 31 "Cybersecurity in AI: balancing innovation and risks" (Segurança cibernética em IA: equilibrando inovação e riscos)** na 19ª reunião anual do **Fórum de Governança da Internet** (15 a 19 de dezembro de 2024).

O documento contou com a colaboração das seguintes pessoas:

- Yury Shelikhov, líder de Cibersegurança e Proteção de Dados, Kaspersky
- Vladislav Tushkanov, gerente do Grupo de Desenvolvimento de Pesquisa, Pesquisa em Tecnologia de Machine Learning, Kaspersky
- Alexey Antonov, cientista de dados líder, Análise de Métodos de Detecção, Kaspersky
- Allison Wylde, membro da equipe, PNAI FGI ONU (interoperabilidade)
- Dr. Melodena Stephens, professor de Governança de Inovação e Tecnologia, Mohammed Bin Rashid School of Government, EAU
- Sergio Mayo Macías, gerente de Programas de Inovação, Instituto Tecnológico de Aragão (ITA), Espanha
- Igor Kumagin, especialista em cibersegurança
- Dmitry Fonarev, gerente sênior de Relações Públicas, Kaspersky

Sumário



Introdução

Objetivo

Com as novas tecnologias digitais, surgem também **novos riscos de cibersegurança e vetores de ataque**. Dessa forma, as empresas devem garantir que a integração da IA esteja protegida dessas ameaças.
O conceito de segurança no desenvolvimento de sistemas de IA foi colocado em primeiro plano em várias iniciativas regulatórias, como a Lei de IA da UE ou o Modelo de Estrutura de Governança de IA Generativa de Singapura, com o objetivo de minimizar os riscos cibernéticos associados. A UE está estabelecendo regulamentações rigorosas de IA com a Lei de IA, que visa garantir transparência, segurança e padrões éticos. Os EUA estão mais focados em desenvolver padrões para o setor e em incentivar a inovação do que no estabelecimento de uma legislação rígida. A China está formulando ativamente padrões e regulamentações que apoiam o desenvolvimento de tecnologias de IA, mas também limitam seu uso em determinadas áreas.

Apesar desse progresso regulatório, ainda existem lacunas importantes entre as estruturas gerais e sua implementação prática em um nível mais técnico. Neste artigo, exploramos os **requisitos básicos de cibersegurança** que devem ser considerados na implementação de sistemas de IA. Esses requisitos devem ser aplicados a uma **gama mais ampla de empresas que dependem de componentes de IA de terceiros** para criar suas próprias soluções.

Para implementar a IA com segurança, as organizações precisam de orientação técnica sobre como desenvolver e implantar a IA em sua infraestrutura. Sem orientação adequada, essa implementação pode acarretar riscos importantes. O objetivo deste documento é fornecer diretrizes para desenvolvedores e administradores de sistemas de IA, MLOps e DevOps de IA e utiliza modelos básicos existentes para criar soluções de IA generalizadas, com ênfase particular em sistemas de IA baseados na nuvem. O artigo aborda os **principais aspectos do desenvolvimento, da implantação e operação de sistemas de IA**, incluindo o projeto, melhores práticas de segurança e integração, sem focar o desenvolvimento de modelos básicos.

A **inteligência artificial (IA)** evoluiu para uma tecnologia crítica para a economia global, tornando-se parte de nossa vida cotidiana. Com a IA, as organizações podem automatizar tarefas de rotina, melhorar o atendimento ao cliente e possibilitar o acesso mais rápido e fácil dos funcionários às informações.

> 50%

Um estudo recente da Kaspersky revelou que mais de 50% das empresas implementaram soluções baseadas em IA em suas infraestruturas*.

33%

pretendem adotar essa tecnologia dentro de dois anos.

* More than half of companies use AI and IoT in their business processes (Mais de metade das empresas usa a IA e a IoT em seus processos de negócios), https://www.kaspersky.com/about/press-releases/more-than-half-of-companies-use-ai-and-iot-in-their-business-processes

As ameaças aos sistemas de IA estão crescendo à medida que essa tecnologia é implantada cada vez mais nas organizações. Ataques cibernéticos afetam todos os estágios do desenvolvimento da IA, dos conjuntos de dados aos algoritmos e as saídas dos modelos.

Visão geral do cenário de ameaças à IA

De acordo com a pesquisa da Kaspersky*, os sistemas de IA enfrentam desafios de segurança únicos e em evolução que colocam a operação dos sistemas em risco. Agentes mal-intencionados exploram vulnerabilidades em dados de treinamento, manipulam modelos para alterar seu comportamento e comprometem a integridade do sistema. Isso destaca a necessidade urgente de segurança abrangente em aplicações de IA.

Problemas com o desenvolvimento de modelos

Ao contrário da programação tradicional, em que é possível compreender e testar o comportamento do código explicitamente, os modelos de machine learning, especialmente os modelos de deep learning, que contêm milhões ou bilhões de parâmetros, são inerentemente complexos e geralmente funcionam como uma "caixa preta". Essa complexidade dificulta a previsão e a interpretação completa do comportamento dos modelos. Consequentemente, o risco de erros não detectados que podem ter consequências graves — como a estabilidade financeira de um banco ou até mesmo a vida de um paciente — aumenta significativamente.

Outro problema é o fato de que os modelos de IA às vezes podem basear suas decisões em **propriedades de dados de entrada irrelevantes ou insignificantes**, em vez de utilizar características relevantes. Por exemplo, modelos de reconhecimento de imagem podem aprender a classificar animais como chitas, leopardos e onças concentrando-se apenas nos padrões de suas manchas em vez de usar sua anatomia geral**. Por exemplo, um modelo classificou incorretamente um sofá malhado como um leopardo porque associou o padrão de manchas ao animal. Essas classificações errôneas podem levar a resultados incorretos em aplicações críticas em áreas como saúde, educação, assistência social, transporte, setor governamental, etc.

Inconsistências entre os dados usados no treinamento e aqueles encontrados durante a implantação podem resultar no baixo desempenho do modelo. Por exemplo, se um modelo for treinado com dados coletados de um tipo de instrumento, mas for aplicado a dados de um dispositivo diferente, ele poderá aprender recursos específicos do dispositivo em vez dos objetos subjacentes que deveria reconhecer. Essa incompatibilidade pode levar a previsões ou classificações imprecisas.

Tanto os modelos de treinamento do zero quanto os modelos básicos de ajuste fino podem enfrentar esses desafios. Embora os conjuntos de dados de ajuste fino sejam menores e mais fáceis de gerenciar, eles também podem introduzir correlações artificiais que fazem com que o modelo resultante não esteja alinhado com seu objetivo.

* AI under Attack (A IA sob ataque), https://content.kaspersky-labs.com/se/media/en/business-security/enterprise/machine-learning-cybersecurity-whitepaper.pdf

** Suddenly, a leopard print sofa appears (De repente, surge um sofá com estampa de leopardo), https://web.archive.org/web/20200208171948/http://rocknrollnerd.github.io/ml/2015/05/27/leopard-sofa.html

Ataques em modelos de IA

Agentes mal-intencionados podem usar vários métodos para atingir modelos de IA. Abaixo estão exemplos de como os invasores exploram vulnerabilidades no projeto, no treinamento e nos mecanismos de interação de IA:

Formas de ataque	Descrição
 Envenenamento de dados: comprometimento da integridade do modelo	O envenenamento de dados envolve a injeção de dados maliciosos no conjunto de dados de treinamento por um invasor com o objetivo de influenciar o comportamento do modelo. Ao criar e adicionar cuidadosamente amostras envenenadas*, os invasores podem fazer com que o modelo tome decisões erradas ou aplique classificações incorretas em certas entradas. Esse tipo de ataque pode comprometer a integridade do modelo e prejudicar sua confiabilidade. Isso também se aplica ao ajuste fino dos modelos básicos.
 Ataques adversários: manipulação invisível da IA	Os ataques adversários envolvem modificações sutis nos dados de entrada que fazem com que o modelo de IA os classifique incorretamente, enquanto as mudanças passam despercebidas pelos humanos**. Os invasores adicionam um ruído especialmente criado às entradas, fazendo com que o modelo produza saídas incorretas enquanto a entrada parece inalterada para observadores humanos.
 Memorização de dados de IA: risco de exposição não intencional	Os modelos de IA mais modernos podem memorizar de modo não intencional determinados detalhes de seus dados de treinamento, especialmente se os dados contiverem amostras únicas ou excepcionais. Os invasores conseguem explorar isso usando técnicas para extrair informações confidenciais que o modelo armazenou inadvertidamente. Dessa forma, poderiam expor dados pessoais do usuário ou informações comerciais confidenciais.
 Injeção de prompt: uma ameaça aos grandes modelos de linguagem	A injeção de prompt é uma ameaça específica de LLMs como o ChatGPT. Os desenvolvedores programam os LLMs para executar tarefas fornecendo instruções iniciais em linguagem natural. Como os usuários também interagem com o modelo usando linguagem natural, o modelo não consegue distinguir inerentemente entre instruções do desenvolvedor e entradas do usuário. Os invasores podem criar entradas que substituem ou manipulam o comportamento do modelo, fazendo com que ele execute ações não intencionais ou divulgue informações confidenciais. Esses prompts maliciosos podem ser inseridos diretamente pelo usuário ou incorporados aos dados processados pelo modelo, como documentos ou páginas da Web.

Esses são apenas os ataques mais relevantes; uma descrição completa de todos os ataques possíveis aos sistemas de IA está além do escopo deste documento.

* Understanding Data Poisoning Attacks (Compreendendo ataques de envenenamento de dados), https://bdtechtalks.com/2020/10/07/machine-learning-data-poisoning/
Attacks Against Machine Learning: An Overview (Ataques contra machine leargning: uma visão geral), (https://elie.net/blog/ai/attacks-against-machine-learning-an-overview/

** Explaining and Harnessing Adversarial Examples (Explicando e aproveitando exemplos de ataques adversários), https://arxiv.org/abs/1412.6572
Adversarial Attacks and Defenses in Deep Learning (Ataques adversários e defesas no deep learning), https://arxiv.org/abs/2201.06192
How to Confuse Antimalware Neural Networks: Adversarial Attacks and Protection (Como confundir redes neurais antimalware: ataques adversários e proteção), https://securelist.com/how-to-confuse-antimalware-neural-networks-adversarial-attacks-and-protection/102949/

Vulnerabilidades de segurança tradicionais

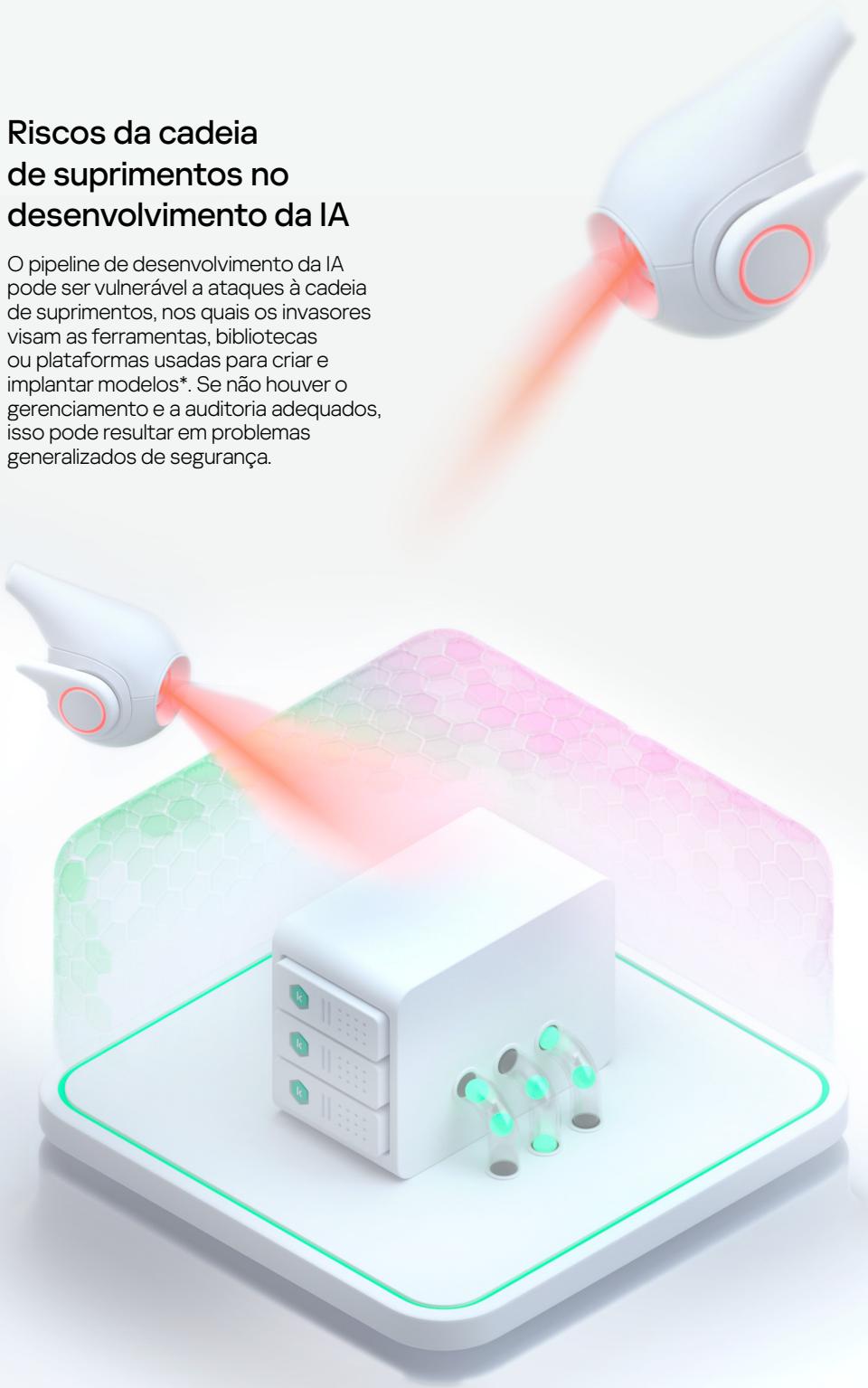
Os modelos de IA podem ser suscetíveis aos pontos fracos da segurança tradicional:

Vulnerabilidades de IA de recursos de terceiros

Os sistemas de IA geralmente dependem de modelos de terceiros ou conjuntos de dados obtidos em repositórios abertos. Esses recursos podem conter erros não intencionais ou backdoors propositais inseridos por agentes mal-intencionados. A incorporação desses componentes comprometidos pode introduzir vulnerabilidades no sistema de IA, afetando sua segurança.

Riscos da cadeia de suprimentos no desenvolvimento da IA

O pipeline de desenvolvimento da IA pode ser vulnerável a ataques à cadeia de suprimentos, nos quais os invasores visam as ferramentas, bibliotecas ou plataformas usadas para criar e implantar modelos*. Se não houver o gerenciamento e a auditoria adequados, isso pode resultar em problemas generalizados de segurança.



Erros no código que expõem vulnerabilidades da IA

Erros no código das interfaces de acesso à IA podem causar vulnerabilidades.

Risco de roubo de componentes da IA

Sem a proteção adequada, conforme mostramos neste documento, é possível roubar os sistemas de IA ou seus componentes críticos, como modelos ou conjuntos de dados.

* Compromised PyTorch-nightly dependency chain between December 25th and December 30th, 2022 (Cadeia de dependência do PyTorch-nightly comprometida entre 25 e 30 de dezembro de 2022), https://pytorch.org/blog/compromised-nightly-dependency

Diretrizes

Conscientização e treinamento em cibersegurança

A implementação de novas tecnologias, como a IA, exige **apoio da liderança, estabelecimento de políticas e governança internas e treinamento especializado para funcionários sobre os riscos e ameaças associados à IA.** Esse treinamento é crítico devido à rápida evolução da tecnologia. Muitos dos recursos de IA disponíveis para desenvolvedores não estão suficientemente maduros ou não adotam totalmente os princípios de segurança por padrão e segurança por design. Como resultado, os desenvolvedores de sistemas de IA têm a sobrecarga de tratar dos riscos potenciais que precisam ser explicados.

Dessa forma, as organizações devem considerar a implementação destas medidas, além das práticas de segurança padrão:

1

A liderança da organização precisa estar ciente dos riscos de segurança associados ao uso de serviços de IA e deve aprender a gerenciá-los.

2

As políticas de segurança da organização devem ser atualizadas para atender aos requisitos específicos dos serviços de IA, garantindo que todos os funcionários e fornecedores estejam familiarizados com elas.

3

A política deve descrever os riscos associados ao desenvolvimento e uso de serviços de IA, bem como as restrições atuais ao seu uso em conformidade com a legislação local.

4

A política deve definir funções e responsabilidades relativas ao uso de serviços de IA dentro da empresa.

5

Deve ser desenvolvido ou adquirido um curso de treinamento corporativo sobre o uso seguro de serviços de IA dentro da empresa, e todos os funcionários e novos contratados devem fazê-lo. O programa deve abranger políticas organizacionais, ameaças existentes e exemplos de incidentes, medidas de proteção contra ameaças, leis aplicáveis e outros tópicos relevantes, além de incluir exercícios de simulação baseados em cenários, se for o caso. Após a conclusão do curso, os funcionários deverão fazer um teste. O curso deve ser atualizado regularmente.

6

Os cursos de segurança da informação existentes devem ser atualizados para incluir novos métodos usados por agentes mal-intencionados que exploram serviços de IA para atacar empresas, como a geração de texto. Exemplos incluem clonagem de voz, manipulação de fotos ou geração de vídeos falsos.

7

O monitoramento da legislação relacionada ao uso seguro de serviços de IA deve ser sistematizado. As políticas internas e os programas de treinamento devem ser atualizados periodicamente.

A modelagem de ameaças ajuda a identificar, entender e mitigar potenciais riscos de segurança nos estágios iniciais do desenvolvimento do sistema de IA.

Modelagem de ameaças/Avaliação de riscos

Esse processo é particularmente importante para sistemas de IA, pois trata-se de uma tecnologia emergente, com riscos que estão em constante evolução e adaptação. Uma avaliação de riscos pode ajudar a prever e preparar-se para esses desafios. Além disso, a modelagem de ameaças pode ajudar desenvolvedores iniciantes a se preparar melhor para os desafios associados ao desenvolvimento de sistemas de IA, e também permitirá que identifiquem e mitiguem proativamente os prontos fracos do sistema de IA antes que sejam explorados.

Para garantir um processo eficaz de modelagem de ameaças, a organização deve seguir estas **recomendações**:

Selecione uma metodologia de avaliação de riscos (por exemplo, STRIDE, DREAD, LINDDUN, PASTA, TRIKE) e desenvolva procedimentos para conduzir avaliações de riscos e modelagem de ameaças para serviços de IA. A metodologia de avaliação também deve incluir uma estrutura para determinar os níveis de risco, estratégias para gerenciar riscos dentro da organização (incluindo limites de risco aceitáveis), procedimentos para monitoramento de riscos e a designação de pessoal responsável para supervisionar o processo.

Ao gerenciar riscos, considere classificar as ameaças nas seguintes categorias:

- ameaças decorrentes da não utilização de serviços de IA,
- ameaças decorrentes de não conformidades,
- ameaças decorrentes do uso indevido de serviços de IA pelos usuários,
- ameaças aos modelos de IA e aos conjuntos de dados usados para treinamento,
- ameaças que os modelos de IA apresentam aos serviços,
- ameaças associadas a dados e
- ameaças ambientais, à sociedade e à governança (ASG).

Deve ser realizada uma avaliação de riscos de todos os serviços de IA existentes e recém-desenvolvidos.

As avaliações de riscos devem incluir a identificação de potenciais agentes de ameaças ou invasores, bem como ameaças e riscos identificados. Devem ser usados materiais de referência como NIST-AI-600-1, MITRE ATLAS, OWASP Top 10 for LLM Applications, DASF e CSA para identificar ameaças e riscos conhecidos.

As informações sobre riscos identificados em serviços de IA devem ser comunicadas à liderança da organização.

Segurança da infraestrutura (nuvem)

Os serviços de IA normalmente são fornecidos como serviços de nuvem e, muitas vezes, exigem infraestrutura especializada. Por exemplo, servidores equipados com GPUs, FPGAs, ASICs ou TPUs. Considerando a confidencialidade dos sistemas de IA, eles devem ser protegidos de acordo com as **estruturas de segurança cibernética mais avançadas**, como o NIST Cybersecurity Framework ou outro com padrões semelhantes. Os serviços de IA costumam usar software de código aberto ou gratuito, como TensorFlow, PyTorch ou Keras, além de bibliotecas como Pandas, NumPy e SciPy. Para proteger esse ambiente, os **seguintes requisitos** devem ser considerados:

1

Identifique todos os ativos e mantenha um inventário de ativos de informação, como conjuntos de dados para treinamento e teste de modelos, dados para treinamento de ajuste fino, bancos de dados, cartões de dados, modelos e riscos, dados de entrada e saída de/para o serviço, pesos e hiperparâmetros de modelos, dados de log de sistemas LLM.

2

Controle o acesso em todos os níveis, incluindo a rede, sistemas operacionais, bancos de dados, software, dados e modelos. Implemente a autenticação de dois fatores (2FA) para acesso administrativo.

3

Registre todos os eventos e garanta que os dados do log estejam protegidos. Monitore incidentes de segurança e possíveis violações.

4

Implemente proteções contra malware e outros tipos de ataques. Aplique regularmente os patches de segurança dos componentes da infraestrutura.

5

Segmente a rede para proteger áreas sigilosas. Use criptografia para dados em trânsito e em repouso.

6

Garanta a integridade dos dados críticos e verifique a autenticidade das bibliotecas e dos modelos em uso.

7

Forneça redundância de servidor e de canais de comunicação. Faça backups regulares e garanta seu funcionamento adequado.

8

Armazene as chaves com segurança em um KeyVault.

9

Aplique os princípios de privilégio mínimo e confiança zero em toda a infraestrutura.

Dependendo da infraestrutura de suporte aos serviços de IA, outros requisitos podem incluir:

Usar um gateway de API para gerenciar o acesso aos modelos e lidar com a autenticação via APIs.

Implementar medidas de segurança específicas para ambientes Kubernetes.

Seguir as práticas recomendadas para proteger serviços baseados na nuvem.

Garantir a integridade do código-fonte, dos dados de treinamento, modelos e scripts de automação.

Isolar dados de treinamento, modelos e ambientes de treinamento para evitar vazamento ou contaminação de dados.

Cadeia de suprimentos e segurança de dados

Os ataques à cadeia de suprimentos representam uma ameaça significativa à infraestrutura de qualquer organização, e a arquitetura de IA não é exceção. Há casos conhecidos em que bibliotecas especializadas para treinamento de redes neurais foram alvo de ataques desse tipo. Entretanto, com a IA, surge uma preocupação específica relacionada à segurança dos provedores de serviços e à proteção dos modelos de machine learning.

O acesso a modelos avançados de IA, especialmente LLMs, geralmente depende de soluções baseadas na nuvem. No entanto, a indisponibilidade de certos modelos em determinadas regiões, juntamente com outras restrições, pode levar os funcionários e desenvolvedores da empresa a optar por serviços de terceiros (proxies) que revendem o acesso a modelos de IA por meio de APIs para realizar suas tarefas do dia a dia. Essa prática introduz riscos significativos, da disponibilização de um vetor adicional para vazamentos de dados, no caso de um incidente de segurança no serviço de proxy, até o uso indevido dos dados obtidos para revenda ou no treinamento de suas próprias versões de LLMs. **É importante entender esses riscos,** analisar cuidadosamente as políticas de privacidade do provedor principal e do proxy e conduzir treinamentos de conscientização em toda a empresa sobre os perigos de usar serviços de terceiros para tarefas de trabalho.

Para mitigar esses riscos, a organização pode optar por implantar um serviço de LLM local. Essa abordagem garante que os dados confidenciais processados pelo LLM permaneçam dentro da empresa, desde que certas condições especificadas (por exemplo, isolamento da rede, desativação da telemetria, etc.) sejam atendidas. Entretanto, além dos riscos associados às vulnerabilidades em estruturas de ML, esse método envolve ameaças relacionadas à existência de backdoors nos modelos. Nesse contexto, isso significa que os formatos de dados usados para distribuir modelos de machine learning podem ter diferentes níveis de segurança, e alguns formatos permitiriam a incorporação de código arbitrário que pode ser executado durante a execução dos modelos. Pesquisas mostraram que há modelos disponíveis em repositórios públicos que, embora limitados em número, são capazes de executar um código específico ao serem carregados. A indisponibilidade de modelos para download em determinadas regiões ou restrições de licenciamento podem promover a utilização de modelos de repositórios de terceiros em vez dos originais. O uso de formatos seguros, como safetensors, resolveria esse problema, mas é necessário conscientizar desenvolvedores e analistas de dados sobre a importância de selecionar fontes confiáveis de modelos e usar formatos seguros para intercambiar pesos de modelos.

- Certifique-se de que os modelos de IA sejam obtidos de fontes confiáveis e legítimas. Evite usar repositórios de terceiros.
- Utilize formatos seguros, como safetensors, para intercambiar pesos de modelos e evitar o risco de execução arbitrária de código.
- Implemente medidas para detectar e responder a ataques à cadeia de suprimentos em componentes relacionados à IA.
- Avalie e revise as políticas de privacidade de serviços e proxies de terceiros usados para acessar modelos de IA para garantir que estejam em conformidade com os padrões de segurança.
- Implante modelos de IA localmente sob condições que garantam a privacidade dos dados, como isolamento de rede e desativação de recursos de telemetria.
- Estabeleça protocolos para a implantação segura de modelos locais a fim de minimizar os riscos associados a possíveis backdoors em modelos de machine learning.
- Atualize e aplique patches regularmente nas estruturas de machine learning para corrigir vulnerabilidades conhecidas.
- Implemente medidas para garantir que os dados confidenciais processados por modelos de IA não saiam da infraestrutura da organização.
- Ao usar APIs de terceiros, conduza auditorias de segurança de acordo com os principais padrões internacionais, como o OWASP API Security Top 10.

Testes e validação

Depois de realizar a avaliação e identificar os riscos, é fundamental entender como se proteger de erros acidentais ou deliberados no treinamento e na aplicação do modelo. Para tanto, a organização deve considerar a implementação das seguintes medidas:

1

Avalie os possíveis danos que podem ser causados por erros acidentais ou deliberados no sistema. Determine o valor dos dados usados para treinar o modelo de IA e dos dados que ele processa.

2

Determine se estruturas, modelos ou conjuntos de dados de código aberto estão sendo usados para construir o sistema de IA.

3

Identifique a potencial base de usuários: acesso de funcionários da empresa, clientes ou do público em geral.

4

Verifique a adesão às práticas recomendadas de ML na construção de modelos. Certifique-se de que os conjuntos de dados sejam particionados corretamente em conjuntos de treinamento, teste e validação com base no funcionamento do modelo.

5

Ao validar modelos de IA e suas métricas (falsos positivos e falsos negativos), verifique se os critérios para divisão do conjunto de dados são apropriados para a natureza dos dados (por exemplo, particionamento cronológico para dados temporais e prevenção de vazamento de dados).

6

Avalie quais recursos o modelo usa para tomar decisões e se eles são consistentes com a intuição de especialistas na área. Use métodos de interpretação de modelos, como vetores SHAP, para entender o processo de tomada de decisões do modelo.

7

Avalie o desempenho real do modelo para garantir que ele entregue os resultados esperados. Monitore o modelo continuamente, pois a distribuição dos dados de entrada pode mudar ao longo do tempo e potencialmente degradar o desempenho do modelo.

8

Adapte o plano de testes para verificar se o modelo é suscetível a vulnerabilidades exclusivas de modelos de machine learning (por exemplo, ataques adversários e envenenamento de dados).

Relatório de vulnerabilidades

A IA é uma área relativamente nova da tecnologia e está evoluindo rapidamente. Apesar dos benefícios significativos, muitos sistemas de IA são suscetíveis a vulnerabilidades específicas.

Uma das principais questões é que alguns sistemas de IA podem conter vulnerabilidades que podem ser exploradas para obter acesso não autorizado aos seus dados. Outro exemplo de vulnerabilidade em sistemas de IA é o viés, quando modelos são treinados com dados não representativos ou que contêm vieses ocultos. Por exemplo, os sistemas de IA podem ser afetados por um viés preconceituoso, quando estereótipos e suposições sociais incorretas se infiltram no conjunto de dados do algoritmo, ou por um viés de medição provocado por dados incompletos. Como resultado, esses sistemas podem chegar a decisões injustas ou discriminatórias, impactando negativamente os usuários e minando a confiança nas tecnologias de IA.

Para tratar dessas questões, é necessário implementar um mecanismo que permita aos usuários **comunicar vulnerabilidades e vieses identificados** em sistemas de IA. Esse mecanismo de comunicação permitirá que as organizações recebam feedback rapidamente e tomem estas medidas:

1. Estabeleça uma política disponível publicamente que defina vulnerabilidades em sistemas de IA e descreva como os usuários podem registrá-las.
2. Forneça métodos seguros para os usuários registrarem vulnerabilidades, como formulários da Web criptografados ou endereços de e-mail dedicados.
3. Defina procedimentos para avaliar, priorizar e corrigir prontamente as vulnerabilidades registradas.
4. Comunique-se com a pessoa que registrou a vulnerabilidade sobre o status e a resolução do problema.
5. Mantenha os usuários informados sobre vulnerabilidades conhecidas e esforços de correção para gerar confiança e demonstrar responsabilidade.
6. Colabore com pesquisadores de segurança por meio de programas de recompensa por bugs. Isso também ajudará você a ficar por dentro das ameaças emergentes e das melhores práticas de segurança de IA.

Defesa contra ataques específicos de ML

Considerando os avanços atuais no desenvolvimento da IA, alguns dos componentes de IA podem ser vulneráveis a ataques específicos de ML. Esses ataques podem explorar vulnerabilidades, por exemplo, alimentando o modelo deliberadamente com dados malformados ou comandos ocultos. Assim, organizações que usam a IA gratuita para desenvolver seus sistemas devem estar cientes desses riscos. A proteção contra ataques específicos de ML requer a **implementação de várias medidas de segurança**, como:

Incorporar exemplos de ataques adversários* no conjunto de dados de treinamento para que o modelo aprenda a lidar com essas entradas de maneira mais eficaz.

Aplicar técnicas de destilação que ajudem a tornar o modelo mais resiliente a entradas adversas, simplificando seu processo de tomada de decisões.

Considerar o uso de modelos monotônicos** que podem melhorar a estabilidade e reduzir a suscetibilidade à manipulação adversária.

Introduzir sistemas capazes de detectar entradas anômalas ou adversas nas solicitações de usuários, permitindo que o modelo detecte e rejeite tentativas maliciosas antes de processar os dados.

Para oferecer proteção contra o envenenamento de dados, os desenvolvedores de sistemas de IA devem **analisar as amostras de treinamento** em busca de objetos anômalos e comparar o desempenho dos novos modelos com versões anteriores para identificar quaisquer mudanças bruscas nas propriedades do modelo.

E, para proteção contra ataques de injeção de prompt nos LLMs, os desenvolvedores de sistemas de IA podem implementar um sistema que analisa solicitações de usuários recebidas ou outros dados de terceiros inseridos na entrada do LLM. Outra abordagem é analisar as respostas a essas solicitações e avaliar se são compatíveis com a tarefa atual do sistema.

* How to confuse antimalware neural networks. Adversarial attacks and protection (Como confundir redes neurais antimalware. Ataques adversários e proteção), https://securelist.com/how-to-confuse-antimalware-neural-networks-adversarial-attacks-and-protection/102949/

** Monotonic models for real-time dynamic malware detection (Modelos monotônicos para detecção dinâmica de malware em tempo real), https://arxiv.org/pdf/1804.03643

Atualizações de segurança e manutenção periódicas

O campo da IA, especialmente a aplicação de LLMs, ainda é relativamente recente, e nem sempre a qualidade do código é de alto padrão. Como resultado, muitas estruturas e ferramentas usadas para trabalhar com machine learning podem conter um **número significativo de vulnerabilidades**. Felizmente, na fase atual, as estruturas populares estão sendo atualizadas ativamente para os padrões de qualidade de produção, com novas versões e atualizações de segurança regulares. Além disso, já há programas de recompensa por bugs para descobrir vulnerabilidades na infraestrutura de ML e ajudar a resolver esses problemas rapidamente.

Isso ressalta a importância de monitorar continuamente o estado da sua infraestrutura, das plataformas usadas para controlar experimentos até as bibliotecas projetadas para se comunicar com serviços de nuvem. Manter a infraestrutura atualizada pode levar mais tempo do que o necessário para projetos em campos com evolução mais lenta. Além disso, as versões mais recentes das bibliotecas podem causar problemas de compatibilidade, exigindo maior investimento no desenvolvimento e na manutenção da funcionalidade do código que depende dessas bibliotecas. Esses custos precisam ser considerados durante o planejamento de iniciativas de IA.

Um outro risco associado ao uso de modelos de IA baseados em nuvem, como os LLMs, é o **ciclo de vida relativamente curto de cada versão do modelo**. O modelo selecionado para um projeto pode ser substituído pelo provedor da plataforma por uma nova versão em um curto período de tempo. Embora se espere que a qualidade geral dos modelos melhore, seu comportamento para tarefas específicas e sua resiliência a ataques podem mudar. Isso pode, por exemplo, incluir injeções de prompt ou tentativas de obter saídas não autorizadas por meio de jailbreaks. Para garantir uma transição tranquila entre os modelos sem comprometer a qualidade das tarefas posteriores ou o nível de segurança, é necessário fazer um planejamento antecipado:

1. Certifique-se de que a infraestrutura esteja sempre atualizada com os patches de segurança e as atualizações de estrutura mais recentes.
2. Participe ativamente de programas de recompensa por bugs e use ferramentas de verificação de vulnerabilidades para detectar pontos fracos nas estruturas de ML e na infraestrutura de IA.
3. Avalie e aplique regularmente atualizações de segurança de ferramentas e bibliotecas de machine learning para reduzir a exposição a vulnerabilidades conhecidas.
4. Planeje-se para possíveis problemas de compatibilidade ao usar as versões mais recentes de bibliotecas e estruturas alocando recursos de desenvolvimento e teste.
5. Implemente uma estratégia para gerenciar o ciclo de vida de modelos de IA baseados em nuvem, com planos de transição para novas versões do modelo à medida que forem lançadas pelo provedor.

Conformidade com padrões internacionais

À medida que observamos o rápido desenvolvimento da regulamentação da IA, a conformidade com as leis relevantes e a adesão às práticas recomendadas tornam-se cada vez mais importantes. Primeiro, os dados de treinamento de IA podem ser coletados de diversas fontes em diferentes jurisdições, o que dificulta o processamento e o uso dessas informações. Além disso, os modelos geralmente são obtidos de **repositórios abertos**, o que acrescenta um nível de incerteza à sua conformidade com os requisitos regulatórios de uma jurisdição específica.

Assim, os desenvolvedores da IA enfrentam a difícil tarefa de garantir a conformidade com todos os requisitos legais nos países em que o sistema será usado. A melhor estratégia nessa situação é **seguir os padrões dos líderes em regulamentação de IA**, como China, União Europeia ou Estados Unidos. Muitos países já estão compartilhando suas abordagens e implementando requisitos semelhantes, permitindo que os desenvolvedores se preparem com antecedência para a implantação global do sistema:

1

Estabeleça diretrizes de uso e desenvolvimento ético da IA para garantir a transparência e a responsabilidade nos processos relacionados.

2

Garanta que todos os dados coletados de fontes distintas estejam em conformidade com as leis de privacidade de dados em cada jurisdição, como o Regulamento Geral de Proteção de Dados (RGPD) na Europa ou a Lei de Privacidade do Consumidor da Califórnia (CCPA) nos EUA.

3

Ao usar modelos de IA de repositórios abertos, verifique se eles estão em conformidade com todos os direitos de propriedade intelectual.

4

Siga as principais estruturas regulatórias, como a Lei de IA da União Europeia ou a Declaração de Direitos de IA dos EUA, pois muitas vezes elas são usadas como referência por outros países.

5

Mantenha-se a par das novidades e desdobramentos nas regulamentações de IA em todo o mundo.

6

Audite regularmente os modelos e sistemas de IA para verificar a conformidade com os padrões internacionais a fim de identificar e mitigar potenciais riscos legais e éticos.

Conclusão

Assim como a maioria das inovações tecnológicas, as tecnologias de IA apresentam grandes oportunidades e ameaças significativas. Os riscos de segurança cibernética associados à IA e seu impacto na sociedade dependem do comportamento e das intenções do desenvolvedor. Para implementar a IA com segurança, as organizações precisam seguir orientações técnicas sobre como desenvolver e implantar a IA em sua infraestrutura, pois não seguir as recomendações adequadas durante esse processo pode representar riscos significativos. É essencial que as organizações estabeleçam uma cultura de segurança e responsabilidade durante todo o ciclo de vida da IA e incorporem controles básicos de segurança, desde a avaliação de riscos e testes de sistemas até a proteção das cadeias de suprimentos e a manutenção contínua. A implementação bem-sucedida dos requisitos apresentados ajudará a **mitigar os riscos** relacionados à introdução de sistemas de IA nas operações da empresa.

Saiba mais

www.kaspersky.com

© 2024 AO Kaspersky Lab.
Marcas registradas e marcas de serviço
pertencem aos seus respectivos proprietários.

#kaspersky
#bringonthefuture