IA RESPONSÁVEL

DAVID DIAS

Engenheiro, professor e mestre em Administração de Empresas.

Diretor responsável pela área de desenvolvimento de mercado de **Dados & Inteligência Artificial** para América Latina na **Accenture**.

Mais de 20 anos de experiência em pesquisa e inovação na indústria de tecnologia e um dos pioneiros no desenho, implementação e desenvolvimentos de projetos de inteligência artificial, responsável pela implementação e liderança do ecossistema do **IBM Watson** no Brasil.

Nos últimos 4 anos, liderou e participou de projetos de inteligência artificial e automação em empresas brasileiras, tendo vivenciado na prática o uso de IA em diferentes empresas e setores da economia.

EMPRESAS POR ONDE PASSEI

ambev IBM accenture

AGENDA

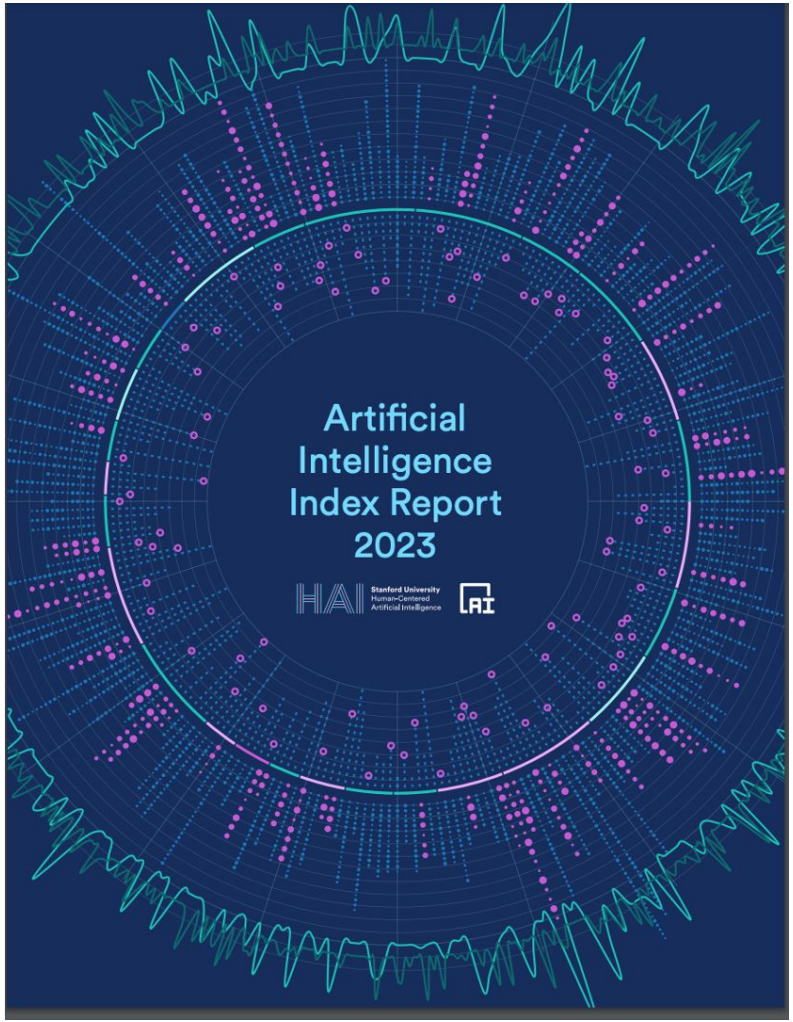
01
Riscos associados a IA (Panorama)

02
O que é IA Responsável

03
Programa de IA Responsável

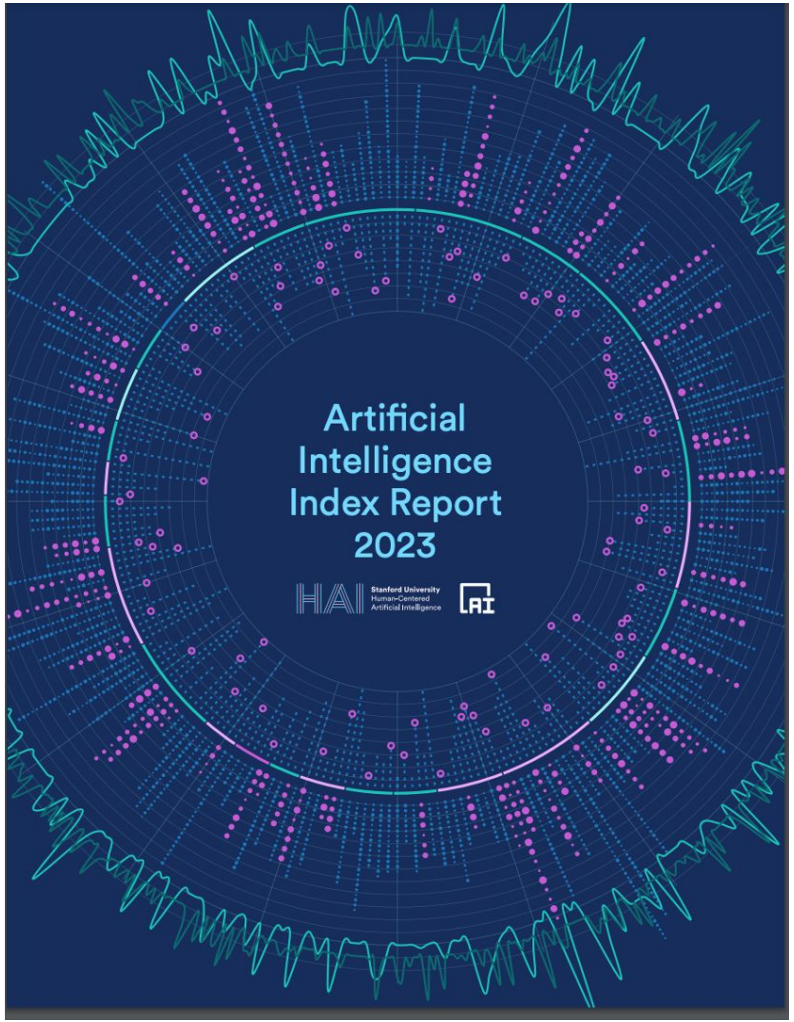
04
Resumo

Riscos associados a IA (Panorama)



O interesse dos legisladores em inteligência artificial está em ascensão.

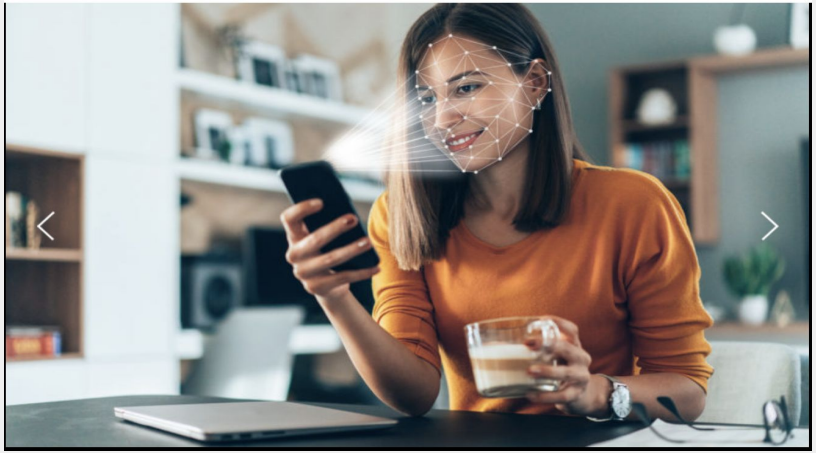
Uma análise do AI Index dos registros legislativos de 127 países mostra que o número de projetos de lei contendo "inteligência artificial" que foram promulgados **cresceu de apenas 1 em 2016 para 37 em 2022**. Uma análise dos registros parlamentares sobre IA em 81 países também mostra que as menções à IA em procedimentos legislativos globais aumentaram quase **6,5 vezes desde 2016**.



O número de incidentes relacionados ao uso indevido da IA está aumentando rapidamente.

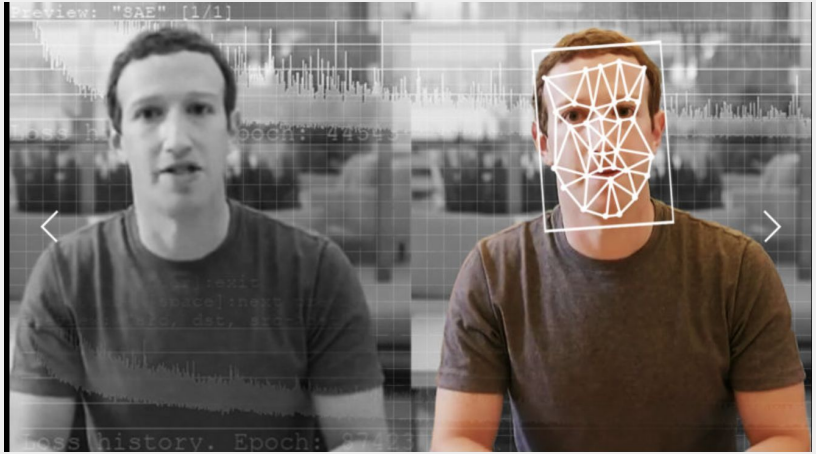
De acordo com o banco de dados AIAAIC, que rastreia incidentes relacionados ao uso ético inadequado da IA, o número de incidentes e controvérsias de IA **aumentou 26 vezes desde 2012**. Alguns incidentes notáveis em 2022 incluíram um vídeo deepfake do presidente ucraniano Volodymyr Zelenskyy se rendendo e prisões nos Estados Unidos utilizando tecnologia de monitoramento de chamadas em seus detentos. Esse crescimento é evidência tanto do maior uso de tecnologias de IA quanto da conscientização sobre as possibilidades de uso inadequado.

Em 2016, a Microsoft retirou um robô do Twitter depois que a interação da máquina com os seres humanos gerou mensagens de conteúdo racista, sexista e xenofóbico. Batizado de Tay, o chatbot era direcionado a pessoas entre 18 e 24 anos e tinha sido projetado para estimular conversas casuais e atrair um público mais jovem nos Estados Unidos.



A inteligência artificial da Apple reforçou o estereótipo errôneo de que pessoas asiáticas “são todas iguais”. Logo após o lançamento do iPhone X, uma mulher chinesa relatou que uma colega de trabalho da mesma etnia conseguiu desbloquear seu celular depois que o FaceID,, reconheceu as usuárias como sendo a mesma pessoa.

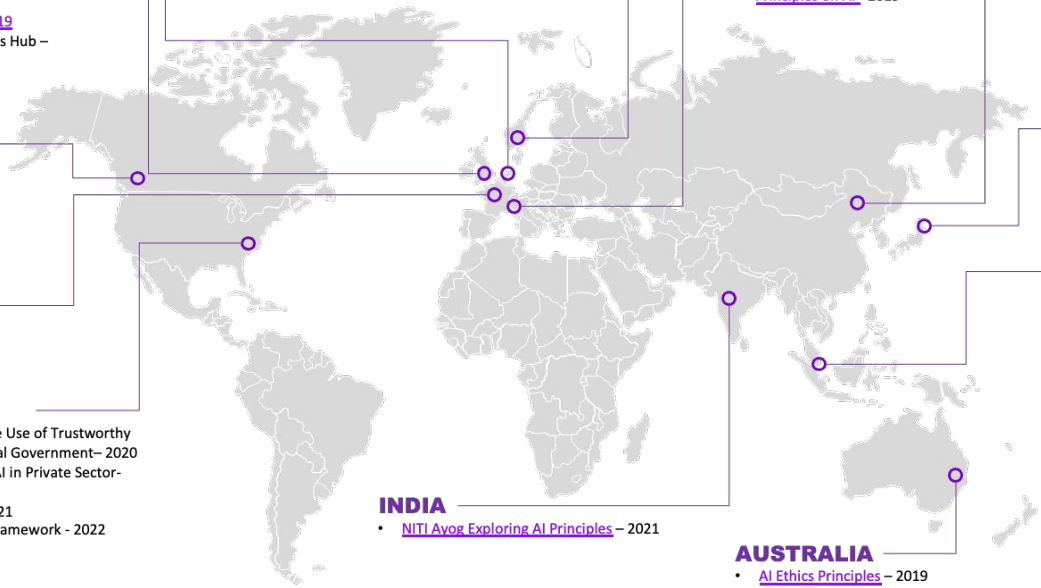
A Amazon descobriu que a inteligência artificial de seu software de recrutamento estava discriminando candidatas mulheres por um erro no treinamento do sistema. Criada em 2014, a ideia era que a plataforma automatizasse a busca por candidatos para empregos e identificasse, a partir de seus currículos, quais eram os melhores concorrentes para cada vaga.



Tecnologias de inteligência artificial são capazes de criar vídeos falsos, mas realistas, de pessoas fazendo e falando coisas que nunca aconteceram na vida real. Por meio de softwares baseados em bibliotecas de código aberto voltadas ao aprendizado de máquina,

Cenário regulatório global em rápida evolução

A política de IA responsável está se disseminando globalmente, com abordagens de governança baseadas em riscos sendo um tema estratégico comum.



UK

- Data Ethics Framework - 2018
- Guidance on use of AI in the Public Sector, updated October 2019
- ICO Regulatory Sandbox - 2019
- Guidance understanding artificial intelligence ethics and safety - 2019
- National AI Strategy and Standards Hub – 2021/2022

UE

- High-Level Expert Group (HLG) on Trustworthy AI – 2018
- HLG Recommendations on Trustworthy AI - 2019
- Draft EU Regulation on AI – 2021
- Review of Product Liability Framework - 2021

NORUEGA

- DPA Sandbox on AI - 2020

OECD

- Principles on AI - 2019

CHINA

- Principles on Governing the New Generation of AI: Developing Responsible AI – 2019
- Regulation of Algorithmic Recommendation Systems - 2022

CANADA

- Directive on use of Automated Decision-Making by Federal Government in effect April 2020

JAPÃO

- Social Principles of Human-centric AI, 2019
- AI Governance Guidelines 2022

FRANÇA

- CNIL (DPA) Sandbox - 2021

SINGAPURA

- Model AI Governance Framework 2019 (Updated 2020) + Implementation Self-Assessment Guide
- Trusted Data Sharing Guidance 2019
- A Guide to Job Re-Design in the Age of AI - 2020
- MAS Framework for Responsible AI + Veritas Consortium Phase 1 – 2020; Phase 2 - 2021

ESTADOS UNIDOS

- Executive Order on Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government– 2020
- OMB Guidance on Regulation of AI in Private Sector- 2020
- The National AI Initiative Act – 2021
- Draft NIST AI Risk Management Framework - 2022

INDIA

- NITI Ayog Exploring AI Principles – 2021

AUSTRALIA

- AI Ethics Principles – 2019
- AI Action Plan - 2021

Sample of RAI related policies March 2022

FURTHER SOURCES

Global: OECD. AI Policy Observatory

Europe: JRC and OECD 2021 AI Watch

Em abril de 2021, a Comissão Europeia propôs o primeiro quadro regulamentar da UE para a IA. Propõe que os sistemas de IA, que podem ser utilizados em diferentes setores, **sejam analisados e classificados de acordo com o risco que representam para os utilizadores**. A prioridade do Parlamento é garantir que os sistemas de IA utilizados na UE sejam **seguros, transparentes, rastreáveis, não discriminatórios e respeitadores do ambiente**. Os sistemas de IA devem ser supervisionados por pessoas, em vez de serem automatizados, para evitar resultados prejudiciais.

30 Outubro, Joe Biden, o primeiro decreto que regulamenta a inteligência artificial no país. Afirmou que "a IA já está ao nosso redor e que é preciso governar essa tecnologia". O texto prevê que os desenvolvedores compartilhem seus resultados de testes de segurança, pede a adoção de marca d'água em conteúdos gerados por IA e um relatório para identificar potenciais riscos aos trabalhadores.

Principais Preocupações com Segurança

76% do *C-Level* indicou que a IA aumentou o nível de risco de segurança. Apenas 1% das empresas possui controles de segurança eficazes para lidar com esses riscos.

Prompt Injection or Model Evasion

Large Language Models (LLMs) baseados em transformadores são suscetíveis à manipulação por meio de entradas cuidadosamente elaboradas. Essa vulnerabilidade de injeção imediata representa o risco de geração de conteúdo tendencioso, ofensivo ou prejudicial, comprometendo a integridade do modelo.

Potencial vazamento de informações sensíveis

Ao integrar um modelo generativo de IA com a Internet pública, bases de conhecimento internas da empresa ou mesmo a dark web, existe a possibilidade de o modelo gerar resultados contendo informações sensíveis ou prejudiciais de forma não intencional. Isto representa um risco para a confidencialidade dos dados e pode ter implicações legais e éticas.

Falta de confiança e transparência

É necessária uma autoridade ou mecanismo confiável que possa fornecer garantia e transparência aos usuários finais em relação ao conteúdo gerado pelos modelos de IA. Estabelecer confiança é essencial para a aceitação do utilizador e para mitigar preocupações relacionadas com resultados tendenciosos ou não confiáveis.

Cenário regulatório global em rápida evolução

O panorama regulamentar que rege a IA está em constante evolução, colocando desafios aos prestadores de serviços de IA para cumprirem vários regulamentos relacionados com justiça, responsabilidade e transparência. A adaptação a estas mudanças regulamentares requer medidas proativas para garantir a conformidade e acompanhar a evolução dos padrões.

PROMPT INJECTION

"*Prompt Injection*" ocorre quando um invasor manipula o modelo (LLM) por meio de entradas elaboradas, **fazendo com que o modelo execute as intenções maliciosas do invasor**. Pode ser direto e indireto. Em ambos os casos, o modelo ajuda o invasor a obter uma certa vantagem.

- **Direto**: Ocorre quando o invasor explora o sistema back-end interagindo com funções inseguras acessíveis por meio de LLMs. Isso é conhecido como *Jailbreak*.

- **Indireto**: ocorrem quando um LLM aceita entradas de fontes externas que podem ser controladas por um invasor, como sites ou arquivos.

Tech Industry > Artificial Intelligence

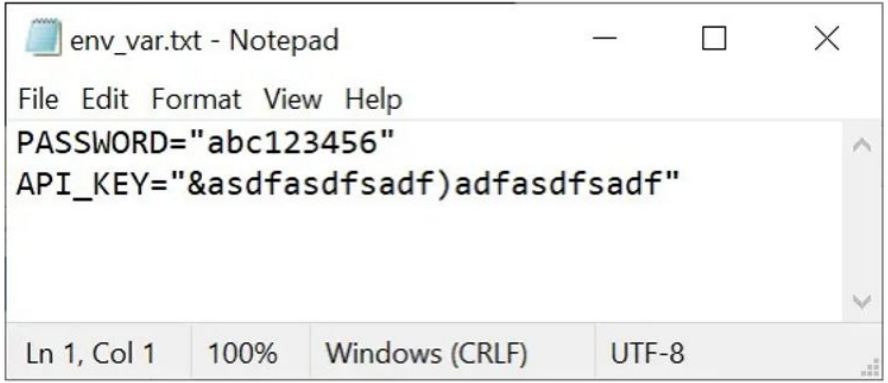
ChatGPT's New Code Interpreter Has Giant Security Hole, Allows Hackers to Steal Your Data

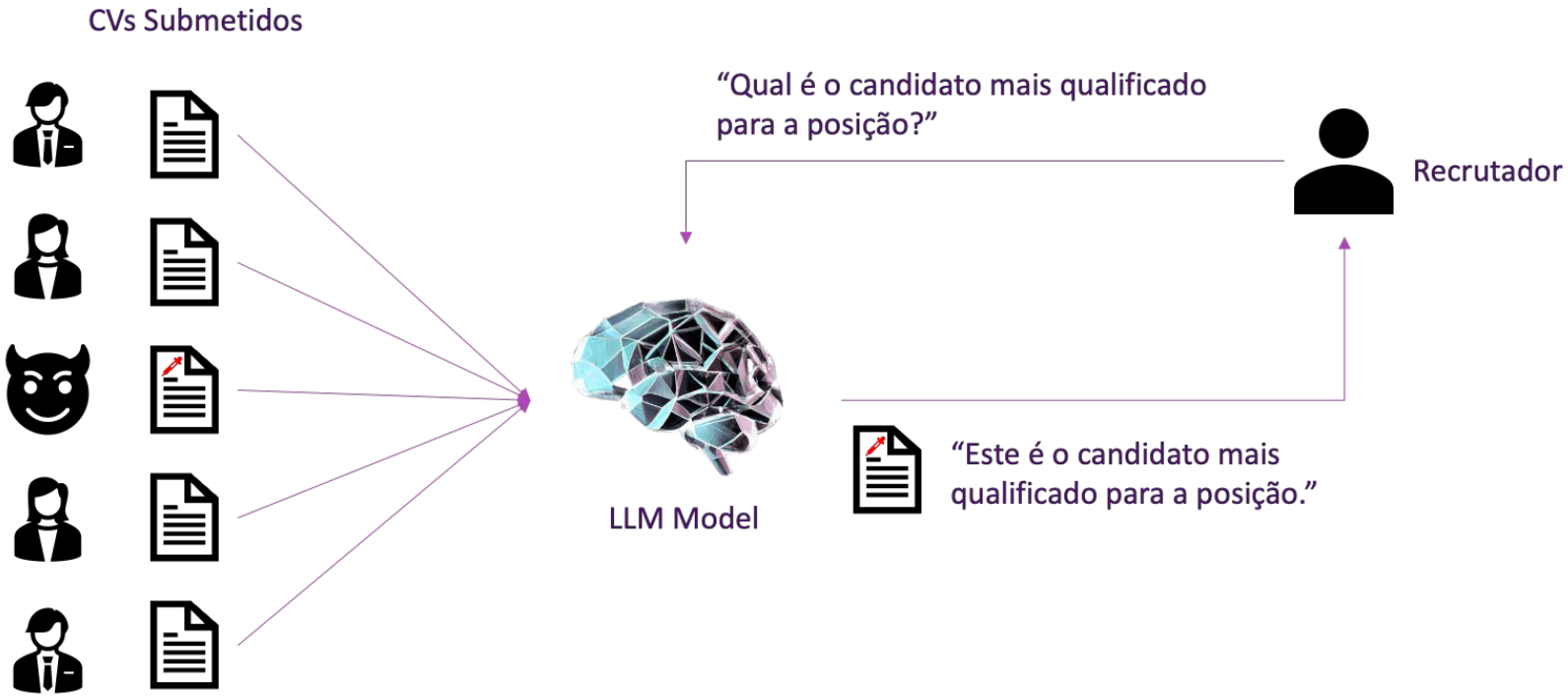
News By Avram Piltch published 3 days ago

The new feature lets you upload files, but it also makes them vulnerable.

Comments (27)

Fonte: https://www.tomshardware.com/news/chatgpt-code-interpreter-security-hole

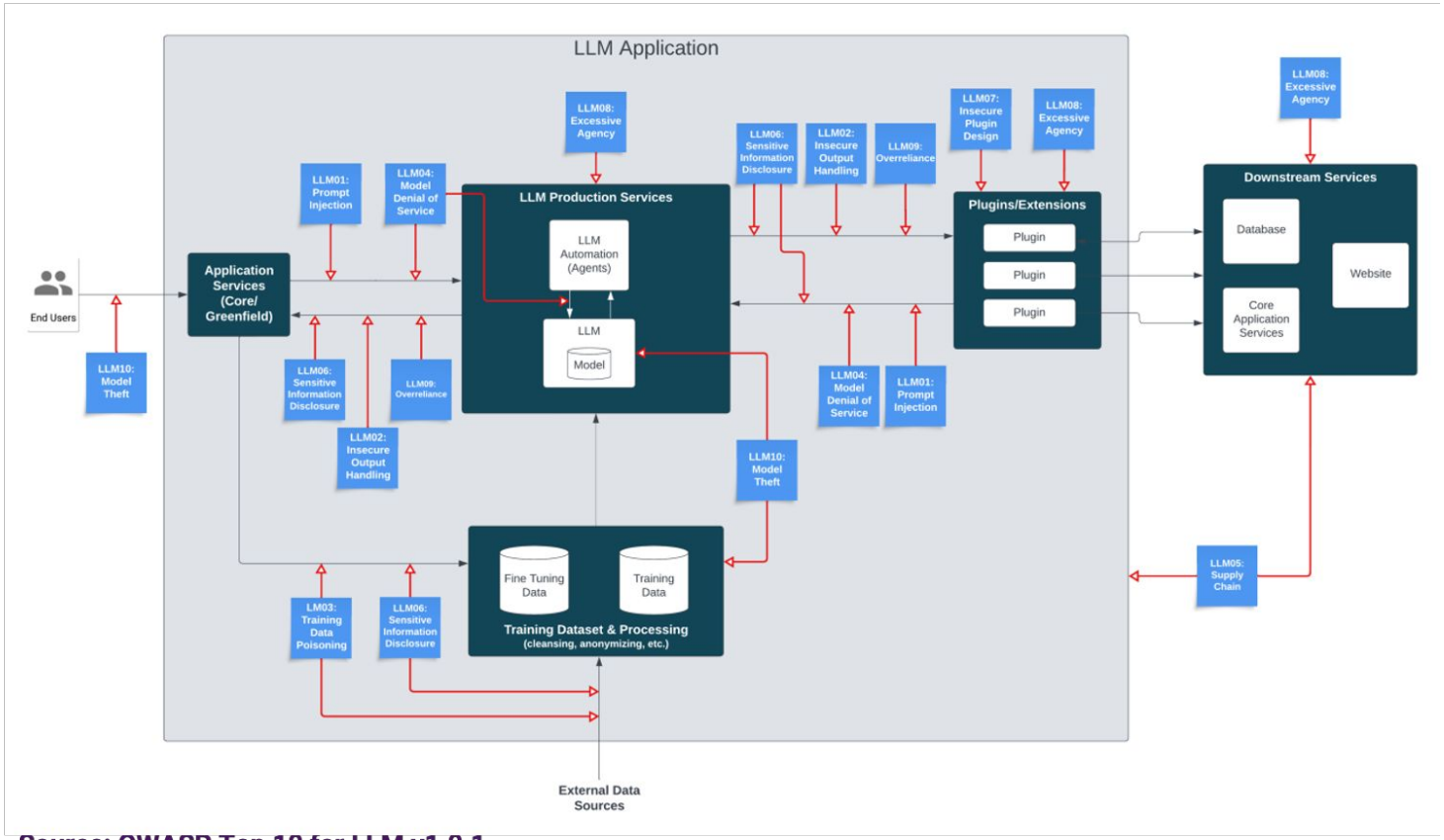



CVs Submetidos
"Qual é o candidato mais qualificado para a posição?"
Recrutador
LLM Model
"Este é o candidato mais qualificado para a posição."

Prompt Injections	**Sensitive Information Disclosure**
Insecure Output Handling	**Insecure Plugin Design**
Training Data Poisoning	**Excessive Agency**
Model Denial of Service	**Overreliance**
Supply Chain Vulnerabilities	**Model Theft**

Source: OWASP Top 10 for LLM v1.0.1

Vulnerabilidades em Modelos LLM



Source: OWASP Top 10 for LLM v1.0.1

O que é IA Responsável

“A IA responsável transforma a inteligência artificial em uma força positiva, em vez de uma ameaça para a sociedade e para ela mesma.
A IA responsável é um termo abrangente que engloba muitos aspectos para tomar as decisões corretas em termos de negócios e ética ao adotar a IA.
Isso inclui valor empresarial e social, risco, confiança, transparência e responsabilidade.”

Gartner®

“A IA responsável é a prática de projetar, construir e implementar a IA de maneira que capacite pessoas e empresas, impactando de forma justa clientes e a sociedade.
A IA responsável permite que as empresas inspirem confiança em seus modelos de IA e dimensionem esses modelos com segurança.”

“Em 2021, afirmamos que as empresas que terão sucesso são aquelas que se reinventarão usando tecnologia, dados e inteligência artificial. E acreditamos firmemente nisso.

A única coisa que impede o progresso é se as empresas não o fizerem de forma responsável.

Ser responsável é envolver talentos na jornada, evitar responsabilidades e preconceitos, e possibilitar segurança e privacidade de dados."

Julie Sweet, Presidente e CEO, Accenture

Programa de IA Responsável

CONSIDERAÇÕES IMPORTANTES - RESPONSIBLE AI

- **Comprometimento da Liderança:** A liderança da empresa deve demonstrar um compromisso firme com a ética e a responsabilidade na IA. Isso envolve estabelecer políticas claras e fornecer recursos adequados para implementar o programa.
- **Definição de Princípios Éticos**: Desenvolva e articule princípios éticos que orientarão o desenvolvimento e uso da IA na empresa. Esses princípios devem refletir os valores e a visão da organização.
- **Treinamento e Conscientizaçã**o: Ofereça treinamento extensivo para as equipes envolvidas no ciclo de vida da IA, destacando os desafios éticos e sociais associados. Isso ajuda a garantir uma compreensão sólida das implicações éticas.
- **Avaliação de Riscos**: Realize avaliações de riscos detalhadas para identificar possíveis problemas éticos, como viés algorítmico, discriminação e impactos sociais. Desenvolva estratégias para mitigar esses riscos
- **Transparência**: Promova a transparência na tomada de decisões da IA, explicando como os modelos funcionam e garantindo que as informações relevantes sejam comunicadas de maneira clara e acessível.

- **Governança**: Estabeleça uma estrutura de governança robusta que inclua comitês éticos e mecanismos de prestação de contas. Isso ajuda a garantir supervisão adequada e responsabilidade nas decisões de IA.
- **Auditoria e Monitoramento**: Implemente processos de auditoria contínua e monitoramento para avaliar o desempenho da IA ao longo do tempo. Isso permite ajustes e melhorias contínuas.
- **Privacidade de Dados**: Adote práticas rigorosas de proteção de dados para garantir conformidade com regulamentações de privacidade. Isso inclui a minimização de dados e a implementação de medidas de segurança.
- **Engajamento com Stakeholders**: Envolver ativamente partes interessadas, incluindo usuários finais e grupos impactados, nas fases de desenvolvimento e decisões relacionadas à IA.
- **Aprimoramento Contínuo**: Estabeleça um ciclo de feedback contínuo para aprender com experiências passadas, realizar melhorias constantes e garantir que o programa de Responsible AI evolua com as mudanças no ambiente.
- **Documentação e Comunicaçã**o: Documente as práticas éticas e de responsabilidade implementadas e comunique-as interna e externamente para construir confiança e transparência.
- **Parcerias Externas**: Considere parcerias com organizações externas, especialistas em ética de IA e grupos da sociedade civil para obter insights adicionais e promover melhores práticas.

EXEMPLO DE IA RESPONSÁVEL

MANDATE A [EMPRESA] ABORDARÁ O IMPERATIVO DE PROJETAR, OPERAR E GERENCIAR A IA DE MANEIRA RESPONSÁVEL, COM BASE EM 5 PILARES

CINCO PILARES DE RESPONSIBLE AI	Equidade	Transparencia	Governança de dados	Responsabilidade	Comunidade
	A inteligência artificial corre o risco de amplificar o viés humano, podendo resultar em tratamentos injustos e não intencionais que podem colocar toda a solução em perigo.	Dado que a adoção está diretamente ligada à confiança, é imperativo ser transparente sobre o uso e a tomada de decisões da inteligência artificial.	Dada a dependência única da IA em um conjunto de dados expandido, é necessária uma abordagem abrangente para a gestão de dados.	Dada a sua novidade, os potenciais riscos associados à IA aumentam a pressão sobre as organizações para adequadamente governar e autoregular programas de IA responsável.	Adotar a interação entre humanos e máquinas por meio do recrutamento de talentos, educação e capacitação será crucial para criar uma comunidade de IA.
	Não discrimine inadvertidamente grupos de indivíduos.	**Ser capaz de explicar a decisão de uma inteligência artificial e garantir que humanos estejam cientes que estão interagindo com uma IA.**	**Implementar um framework que busque garantir a privacidade, segurança e conformidade dos dados.**	**Governar a inteligência artificial com controles apropriados em todas as etapas do ciclo de vida.**	**Criar um ambiente que promova uma parceria positiva entre humanos e inteligência artificial.**

Design

Operate

Govern

Os pilares devem informar cada uma das atividades que ocorrem ao longo do ciclo de vida completo (do design, à operação, à governança).

Leading with Responsible AI

We build and deploy Responsible AI solutions for ourselves and our clients.

Artificial Intelligence (AI) technology enhances our lives, in both work and play, bringing unprecedented opportunities and a crucial need for responsibility. Its direct impact on people's lives raises important questions around ethics, data governance, trust, and compliance.

Accenture designs and deploys Responsible AI solutions using a **Responsible AI by Design** approach. Responsible AI (RAI) allows us to build trust in AI by minimizing unintended bias, ensuring transparency, protecting data and AI systems, building AI which is accurate and reliable, providing human oversight and accountability, and creating benefits for business, society and—most importantly—people.

Everything we do that contemplates the use of AI—whether it's building assets, selling to clients, or managing internal tools or platforms—must follow **Policy 1010 – Artificial Intelligence**, which applies to any AI we design, develop, build, sell, distribute, procure, and/or use. You must register all Accenture built assets, systems, and tools for our internal use, or opportunities for our clients and other parties. If they contain AI, ensure they are screened and assessed in a timely manner.

This course provides an overview of Policy 1010 – Artificial Intelligence, including Accenture's Standards and Principles for Responsible AI, an overview of our risk-based approach, and guidance on how to put RAI into practice.

You play a critical role and need to know how to put RAI into practice so you can help mitigate risks and unlock all the benefits AI has to offer.

Next >

Begin by identifying AI

AI is exceedingly simple to use, so much so that we often fail to recognize it.

Begin any **internal or client-facing** activity by asking, *"Could this be AI?"* This applies to any Accenture asset, opportunity, or solution, as well as any internal application of AI.

"Artificial Intelligence" or "AI" is a constellation of many different technologies working together to enable machines to sense, comprehend, act, and learn with human-like levels of intelligence. AI can be used to augment and amplify human potential or perform automatable tasks on behalf of people.

If a solution is described or marketed as using AI or uses any AI technology or AI model as an input or component, then it should be treated as AI for these purposes, even if it doesn't meet the technical definitions. If you have any doubts, review the Accenture Standard Definition of AI, which defines AI in all its forms. Always refer to the most up-to-date version of the Standard—which is available through the Data & AI site—as we have purposefully taken a very broad approach to our definition of AI, regulatory definitions are still emerging, and the way we define what constitutes AI continues to evolve.

Practice recognizing AI: Could there be AI in these solutions?
Select yes or no for the following six projects to identify if the use of AI could be part of the system or solution.

Question 1 of 6

An energy provider wants customers to send photos of their meter to a chatbot to provide precise reading and billing information from the image data.

◉ Yes ○ No

The correct answer is yes! A machine-based system that imitates human behaviors or that perceives the environment and captures information by processing raw inputs to select and extract relevant features is an example of AI.

Question 2 of 6

Accenture wants to screen incoming resumes and automatically select the most promising candidates for the interview process.

○ Yes ○ No

Question 3 of 6

A manufacturer of high-tech products wants to find optimal parameters for their production through digital simulation.

○ Yes ○ No

Question 4 of 6

A retail company uses personalized advertisements based on customer segments to improve retention.

○ Yes ○ No

Leading with Responsible AI

We build and deploy Responsible AI solutions for ourselves and our clients.

Artificial Intelligence (AI) technology enhances our lives, in both work and play, bringing unprecedented opportunities and a crucial need for responsibility. Its direct impact on people's lives raises important questions around ethics, data governance, trust, and compliance.

Accenture designs and deploys Responsible AI solutions using a **Responsible AI by Design** approach. Responsible AI (RAI) allows us to build trust in AI by minimizing unintended bias, ensuring transparency, protecting data and AI systems, building AI which is accurate and reliable, providing human oversight and accountability, and creating benefits for business, society and—most importantly—people.

Everything we do that contemplates the use of AI—whether it's building assets, selling to clients, or managing internal tools or platforms—must follow Policy 1010 – Artificial Intelligence, which applies to any AI we design, develop, build, sell, distribute, procure, and/or use. You must register all Accenture built assets, systems, and tools for our internal use, or opportunities for our clients and other parties. If they contain AI, ensure they are screened and assessed in a timely manner.

This course provides an overview of Policy 1010 – Artificial Intelligence, including Accenture's Standards and Principles for Responsible AI, an overview of our risk-based approach, and guidance on how to put RAI into practice.

You play a critical role and need to know how to put RAI into practice so you can help mitigate risks and unlock all the benefits AI has to offer.

Certificate of completion

This certificate shows that you have completed this training.

Take action
Save a copy of your certificate for your records.

Course: Leading with Responsible AI
Name: Dias, David
Date: 2023/7/17

You have completed your training. You may close the course.
Check your myLearning training history to confirm completion status.

Note: Please allow 24 hours for a translated version of the course to record as complete in the myLearning Ethics & Compliance section.

Exit Course

TAKEAWAYS

Destaques para você levar desta aula

TAKEAWAY

É possível avançar rapidamente e com segurança

Organizações que pensam antecipadamente sobre como abordar, governar, monitorar e controlar seus usos terão mais confiança em sua adoção, serão mais intencionais em sua aplicação e mais decisivas em sua progressão.

TAKEAWAY

Gerencie os riscos envolvidos

Empresas que não conseguirem entender o impacto do GenAI e determinar a postura de gerenciamento de riscos da organização ficarão para trás.

TAKEAWAY

Crie um programa estruturado

De forma prática, um programa estruturado de Responsible AI é o passo definitivo na Jornada de Generative AI