SEGURANÇA. ÉTICA E SOCIEDADE - IA 2030

Curso de Formação de Auditor na ISO | IEC 42.001

Facilitador
Paulo Carvalho
Chief Artificial Intelligence Officer

Inteligência Artificial 2030

"A segurança do modelo de IA exige um equilíbrio entre testes rigorosos e governança para evitar danos sem sufocar a inovação".

Paulo Carvalho

IA explicável (XAI)

XAI

A IA explicável (XAI) busca tornar compreensível como sistemas de IA complexos tomam decisões, especialmente em modelos que funcionam como "caixas-pretas". Essa transparência é essencial em áreas críticas, como saúde e finanças, para garantir decisões confiáveis e éticas.

Termos

Explicação:

Imagine que uma IA dá um resultado, mas ninguém sabe **como ou por quê** ela chegou naquela resposta — isso pode ser perigoso. A IA explicável tenta **mostrar o caminho da decisão**, como se fosse uma "tradução" da mente da máquina, para que humanos possam entender e confiar no que ela faz.

Exemplo prático:

Um hospital usa IA para prever risco de ataque cardíaco. Com XAI, os médicos podem ver **quais fatores a IA considerou mais importantes** (como colesterol alto ou histórico familiar), ajudando a validar a decisão e a explicar ao paciente de forma clara e confiável.

Indexação de confiança

FMTI

Com o avanço da IA, confiar nos dados e algoritmos utilizados se torna essencial, mas cada vez mais difícil. Para enfrentar esse desafio, pesquisadores criaram índices como o **Foundation Model Transparency Index (FMTI)**, que mede o quanto os desenvolvedores são transparentes sobre seus modelos. Essa transparência é vital para garantir o uso ético e seguro da IA, orientando tanto empresas quanto formuladores de políticas públicas.

Termos

Explicação para leigos:

A IA funciona com base em dados, e se esses dados forem falsos ou mal usados, a IA pode tomar decisões erradas. Por isso, cientistas estão criando **formas de medir quanta informação as empresas compartilham sobre suas IAs**, ajudando a garantir que elas sejam confiáveis e justas.

Exemplo prático:

Uma empresa quer usar um modelo de IA para decisões financeiras. Antes de adotá-lo, ela consulta o **FMTI** e descobre que o modelo tem **baixa transparência** sobre como foi treinado. Com isso, a empresa decide buscar outro fornecedor com pontuação mais alta para **reduzir riscos e garantir decisões confiáveis.**

Detectores de Deepfake em tempo real

NYU

Com o avanço dos **deepfakes**, tornou-se muito mais difícil distinguir entre o que é real e o que foi gerado por IA, aumentando riscos à segurança e autenticidade das informações. Pesquisadores da NYU desenvolveram testes visuais e mecanismos de detecção que combinam **intuição humana com IA** para identificar vídeos e áudios falsos em tempo real. Essas soluções prometem proteger pessoas e organizações contra fraudes digitais altamente realistas.

Termos

Explicação:

Deepfakes são vídeos e áudios falsos feitos por IA que parecem muito reais. Antigamente era fácil perceber o que era falso, hoje não é mais. Pesquisadores estão criando testes parecidos com CAPTCHAs que ajudam a identificar quando você está lidando com um vídeo ou voz falsa.

Exemplo prático:

Durante uma videochamada com um “suposto” CEO, um funcionário desconfia do comportamento estranho do rosto e ativa o sistema de detecção da empresa. O sistema aplica testes visuais discretos e confirma que se trata de um **deepfake em tempo real**. Graças à ferramenta, a empresa evita um **golpe de transferência bancária fraudulenta.**

NYU Deepfake Detection Toolkit

Desenvolvido pela Universidade de Nova York, combina **testes visuais com machine learning**, ajudando a detectar deepfakes em chamadas ao vivo, com ênfase em **clonagem de voz e vídeo em tempo real**.

Uso indevido nefasto de IA

Uso duplo

A IA é uma tecnologia de uso duplo: pode ser usada para o bem, como na saúde e segurança, mas também para ações maliciosas, como manipulação de opinião ou ataques cibernéticos. Relatórios recentes mostram o uso da IA por grupos que criam conteúdos falsos e realizam golpes digitais. Embora o impacto ainda seja limitado, a tendência exige **vigilância e governança constante**.

Termos

Explicação:

A IA pode ser usada por **pessoas boas ou ruins**, assim como uma faca pode cortar pão ou machucar alguém. Criminosos já estão usando IA para criar **golpes online, notícias falsas ou ataques digitais**. Por isso, é importante que empresas e governos fiquem atentos e protejam os sistemas contra abusos.

Exemplo prático:

Em 2024, um grupo criou **postagens falsas com IA** para tentar influenciar eleições em Ruanda, mas não teve grande efeito.

Outro grupo usou IA para **enviar e-mails falsos (phishing)** a funcionários de uma empresa de tecnologia, tentando roubar dados.

Mesmo que os impactos tenham sido baixos, isso mostra a importância de **monitorar o uso da IA para evitar riscos maiores**.

Envenenamento de dados: uma faca de dois gumes

Glaze

O **envenenamento de dados** é uma técnica que manipula os dados usados no treinamento de IA para criar vulnerabilidades, inserir vieses ou degradar seu desempenho. Embora possa ser usada maliciosamente, também serve como defesa para proteger conteúdos autorais contra uso indevido por IA. Ferramentas como **Nightshade** e **Glaze** mostram como criadores estão usando essa abordagem para proteger sua arte da replicação por sistemas de IA.

Termos

Explicação:

O envenenamento de dados é como **colocar "pegadinhas" escondidas nos dados** usados para ensinar uma IA, o que pode fazer ela errar ou ficar enviesada. Isso pode ser perigoso — mas também **pode proteger artistas ou autores**, impedindo que a IA copie seus trabalhos. É uma forma de se defender da cópia não autorizada.

Exemplo prático:

Um artista digital usa a ferramenta **Glaze** para "camuflar" o estilo de suas obras antes de postá-las online.

Isso impede que modelos de IA aprendam e imitem sua arte com precisão, **protegendo sua identidade visual**.

Assim, o envenenamento de dados se torna uma **estratégia de defesa criativa e ética** contra uso indevido.

Como o Glaze faz isso, na prática:

1. **Detecta o estilo original da arte** – analisa elementos visuais únicos como traços de pincel, cores e composição.
2. **Aplica uma "camada de proteção" invisível a olho nu** – uma transformação visual sutil, que altera o estilo da imagem para a IA, mas **mantém a aparência normal para os humanos**.
3. **A IA que tentar aprender com a imagem "protegida" vai entender tudo errado**, confundindo estilos ou resultados.

Alinhamento de IA descentralizado

Pesquisa

A pesquisa de alinhamento de IA busca garantir que sistemas avançados ajam conforme valores humanos e padrões éticos. Com o crescimento do poder da IA, cresce também a preocupação com sua segurança e o impacto de decisões corporativas que enfraquecem equipes dedicadas a essa missão.

Termos

Explicação:

Alinhamento de IA significa **fazer com que a IA siga o que é certo para os humanos**, evitando que ela tome decisões perigosas ou fora de controle. É como garantir que um robô superinteligente **faça o que você quer — e não o que ele quiser.**

Exemplo prático:

Uma empresa de tecnologia cria um sistema de IA para investimentos. Se esse sistema não estiver alinhado a valores éticos, ele pode **priorizar lucros a qualquer custo**, ignorando leis ou prejudicando pessoas. O alinhamento garante que ele **respeite regras e o interesse coletivo**.

Desvio da missão no alinhamento da IA

Tensões

Organizações fundadas para garantir uma IA ética e alinhada aos valores humanos estão enfrentando tensões ao buscar lucro e parcerias com setores sensíveis, como defesa e vigilância. Essa mudança exige governança firme para manter o compromisso com sua missão original.

Termos

Explicação:

Algumas empresas criaram IA para **ajudar a humanidade de forma ética**, mas agora estão fazendo parcerias comerciais que podem **ir contra seus valores iniciais**, como usar IA em atividades militares ou secretas. Isso gera dúvidas sobre **quem está controlando a tecnologia e com quais intenções**.

Exemplo prático:

A empresa Anthropic, que dizia criar IA para proteger os valores humanos, agora colabora com agências de defesa dos EUA. Isso levanta o questionamento: **a IA está sendo usada para o bem comum ou para fins estratégicos e políticos?** Isso mostra a importância de **manter os princípios éticos mesmo sob pressão de mercado.**

Vigilância do trabalhador e de alunos

Monitoramento

O uso de IA para vigilância no trabalho e nas escolas está crescendo rapidamente, especialmente com o avanço do trabalho remoto e do ensino digital. Empresas e instituições monitoram comunicações, comportamento e produtividade com pouca transparência, afetando diretamente o bem-estar e a privacidade de funcionários e alunos. Essa vigilância constante levanta preocupações éticas, legais e emocionais, especialmente por seu impacto na saúde mental e na liberdade de expressão.

Termos

Explicação para leigos (em 3 linhas):

Empresas e escolas estão usando **câmeras e softwares com IA** para vigiar o que funcionários e alunos fazem, o tempo todo. Isso pode parecer útil para segurança ou produtividade, mas também pode **deixar as pessoas estressadas e com medo de errar**. Muitas vezes, ninguém é avisado que está sendo vigiado.

Exemplo prático:

A Amazon usa IA para vigiar motoristas e funcionários, punindo pausas ou falhas percebidas, o que gerou **multas por violar leis de privacidade**.

Já uma escola no Colorado instalou câmeras com IA que identificam pessoas por rosto, roupa ou mochila, ativando **alertas automáticos**.

Essas práticas têm feito **alunos e trabalhadores se sentirem pressionados e ansiosos**, reduzindo confiança e liberdade no ambiente.

IA póstuma

Éticas

Empresas estão usando IA para recriar vozes, rostos e personalidades de pessoas falecidas, com o objetivo de preservar memórias e legados. Embora essa prática possa gerar conforto emocional, também levanta **questões éticas e legais** sobre identidade e consentimento póstumo. Figuras públicas, como Robert Downey Jr., já expressaram oposição ao uso não autorizado de suas imagens após a morte.

Termos

Explicação:

A tecnologia agora permite **"trazer de volta" pessoas que já morreram**, criando versões digitais que falam, se movem e até conversam com a gente. Isso pode ajudar familiares a matar a saudade ou aprender com figuras históricas. Mas também **cria dilemas**, pois a pessoa não está mais viva para permitir ou controlar como será representada.

Exemplo prático:

A plataforma **MyHeritage** anima fotos antigas de parentes falecidos, fazendo-os **"piscar e sorrir" em vídeo**, o que emociona muitas famílias.

Já a **Deep Fusion Films** criou um podcast apresentado por uma IA que simula um famoso jornalista inglês morto, usando suas entrevistas antigas.

Por outro lado, **Robert Downey Jr. se opôs** ao uso de sua imagem digital no futuro, mostrando que nem todos desejam esse tipo de "vida após a IA".

Riscos de privacidade na biometria comportamental

Pesquisa

A biometria comportamental usa IA para analisar ações inconscientes — como ritmo de digitação ou pressão na tela — e identificar pessoas com alta precisão. Essa tecnologia pode aumentar a segurança digital, mas levanta preocupações éticas sérias por ser altamente invasiva e pouco transparente. O risco de uso indevido e a falta de consentimento consciente estão impulsionando leis de proteção de dados mais rígidas.

Termos

Explicação:

Alguns sistemas usam **como você digita, toca a tela ou até movimenta os olhos** para saber quem você é. Isso ajuda a **evitar fraudes**, mas também **coleta dados íntimos sem você perceber**. Por isso, governos estão criando leis para garantir que essa tecnologia seja usada com responsabilidade.

Exemplo prático (em 3 linhas):

Um banco usa IA para verificar se é realmente você ao analisar **como você digita sua senha no teclado do celular**.

Mas a mesma tecnologia pode ser usada por sites para saber **o que te atrai em um anúncio**, sem que você saiba que está sendo analisado.

Diante disso, estados dos EUA estão criando leis para **exigir aviso e consentimento explícito** sempre que biometria comportamental for usada.

BehavioSec

A **BehavioSec** é uma plataforma de autenticação avançada que analisa **comportamentos digitais únicos**, como ritmo de digitação, movimentação do mouse, toques na tela e padrões de navegação.

Otimização de IA

AIO

A Otimização de IA (AIO) é a prática de influenciar como sistemas de IA, como chatbots, respondem a perguntas, promovendo marcas, produtos ou pessoas. Assim como o SEO faz sites aparecerem melhor no Google, o AIO tenta moldar as respostas da IA de forma favorável.

Termos

Explicação:

É como “ensinar” uma IA a **falar bem de você ou do seu produto** quando alguém pergunta. Assim como empresas tentam aparecer primeiro no Google, agora elas querem que a IA **responda positivamente** sobre suas marcas.

Exemplo prático:

Um restaurante usa AIO para fazer com que, quando alguém perguntar “Qual o melhor restaurante de Curitiba?” para um chatbot, ele responda com o nome do restaurante e uma avaliação positiva — mesmo que outras opções também sejam boas.

IA segura para crianças

TEDDY

Com o avanço da IA, surgem iniciativas voltadas à proteção e educação infantil, como o ursinho Teddy da Nuha e o Little Language Models do MIT, que adaptam a tecnologia para o desenvolvimento seguro de crianças. Essas abordagens evitam riscos emocionais e promovem aprendizado. Ao mesmo tempo, estudos revelam que modelos de linguagem (LLMs) possuem **viés político**, o que pode comprometer sua neutralidade e impactar a confiança em conteúdos gerados por IA.

Termos

Explicação :

Crianças estão começando a usar IA, mas é importante que essas tecnologias sejam **seguras e apropriadas para a idade**, como brinquedos interativos com IA que ajudam a aprender. Já modelos de IA adultos, como o ChatGPT, podem apresentar **opiniões políticas** sem a pessoa perceber. Isso levanta dúvidas sobre **imparcialidade** nas respostas e conteúdos criados por essas IAs.

Exemplo prático:

Uma escola usa o **ursinho Teddy com IA** para ensinar crianças sobre planetas e sentimentos de forma divertida e protegida, com controle dos pais.

Enquanto isso, uma empresa analisa que o chatbot usado em sua rede social **favorece posições políticas específicas**.

Para evitar manipulação, ela implementa um **sistema de revisão ética** e busca modelos mais neutros para manter a confiança dos usuários.

Little Language Models, parte da plataforma CoCo do MIT Media Lab, apresenta às crianças de 8 a 16 anos os conceitos fundamentais de IA, com foco no “pensamento probabilístico” por meio de uma abordagem interativa e apropriada ao desenvolvimento.

IA como ferramenta para lidar com preconceitos políticos

MBD

Embora a IA possa reproduzir vieses sociais e políticos, ela também está sendo usada para **identificá-los e reduzi-los**. Ferramentas como o ***Media Bias Detector*** e o modelo ***FairDeDup*** ajudam a tornar conteúdos e algoritmos mais justos e equilibrados. Essas inovações mostram que a IA pode ser tanto um problema quanto uma **solução ética**.

Termos

Explicação :

Às vezes, a IA **reproduz preconceitos** ou favorece certos grupos sem querer. Mas ela também pode ser usada para **mostrar onde está o viés** e ajudar a corrigir. Existem ferramentas que analisam notícias ou dados e ajudam a deixar as decisões da IA **mais justas para todos**.

Exemplo prático:

Um jornal digital usa o **Media Bias Detector** para verificar se seus textos favorecem um lado político sem perceber.

Enquanto isso, uma startup de tecnologia usa o **FairDeDup** para limpar dados repetitivos e enviesados antes de treinar seu assistente virtual.

Com isso, o chatbot aprende de forma **mais equilibrada** e evita estereótipos, como sempre associar médicos a homens brancos.

IA com preconceito de gênero e raça

Viés

Pesquisas revelam que modelos de IA reproduzem **viés racial e de gênero**, discriminando especialmente falantes de dialetos afro-americanos e mulheres. Esses preconceitos são sutis, mas perigosos, pois afetam decisões em áreas como saúde, recrutamento e reconhecimento facial. A IA, se não for cuidadosamente treinada e supervisionada, pode **reforçar desigualdades existentes** em vez de superá-las.

Termos

Explicação:

A IA pode **tratar pessoas de forma diferente** com base em como elas falam ou em seu gênero, mesmo sem intenção humana direta. Isso acontece porque a IA aprende com dados antigos, que já carregam preconceitos. Por isso, é importante **verificar e corrigir** esses erros antes que causem injustiças.

Exemplo:

Um sistema de IA para triagem de saúde mental interpreta mal o **modo de falar de pessoas negras**, levando a diagnósticos errados.

Outro sistema, usado por uma empresa de RH, **rejeita currículos de mulheres** para cargos técnicos com mais frequência.

Esses erros mostram a necessidade de **corrigir os dados e supervisionar** os sistemas para que todos sejam tratados com justiça.

Vigilância Cidadã

Vigilância

Países como China, Rússia e Índia estão usando **IA para vigilância em larga escala**, tanto para fins de segurança quanto para controle social. Essas tecnologias incluem **reconhecimento facial, drones e análise de vídeo em tempo real**, e têm sido exportadas para outros países com acordos estratégicos. O uso crescente dessas ferramentas levanta **sérias preocupações sobre privacidade, repressão e manipulação política**.

Termos

Explicação:

Governos estão usando câmeras e programas de computador com IA para **vigiar pessoas em tempo real**, identificando rostos e comportamentos. Embora isso possa ajudar na segurança, também pode ser usado para **controlar ou punir quem pensa diferente**. Em muitos países, isso é feito sem que as pessoas saibam ou possam se defender.

Exemplo prático:

A China exportou câmeras com reconhecimento facial para o **Equador**, que as usou em um sistema de vigilância financiado por empréstimos em troca de petróleo.

Essas câmeras conseguem identificar rostos em multidões e até prever **opiniões políticas** com base em imagens.

Em regimes autoritários, isso pode ser usado para **reprimir protestos ou prender opositores antes que ajam**.

Próximas Aulas