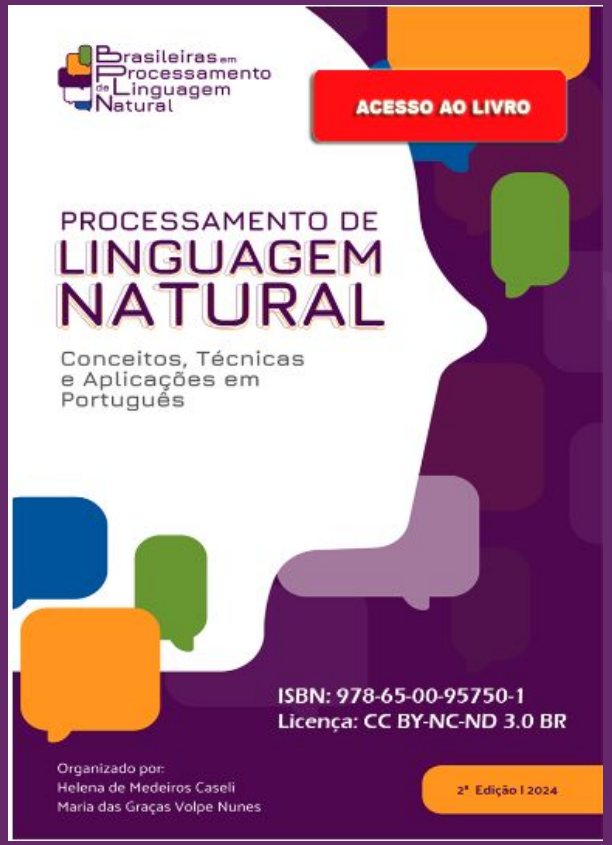
Questões Éticas em IA e PLN

Profa. D.Sc. Mariza Ferro
Instituto de Computação – IC
Universidade Federal Fluminense (UFF)
Núcleo de Referência em IA Ética e Confiável



https://brasileiraspln.com/livro-pln/

Sumário

imagem

Introdução

Conceitos Básicos

Ciência da Computação

Inteligência Artificial

"A ciência e a engenharia de produzir máquinas inteligentes" John McCartthy – 1956

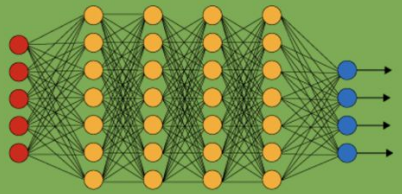
Sistemas que **apresentam comportamento inteligente**, percebendo, raciocinando, agindo e se adaptando, com algum grau de autonomia, para atingir objetivos específicos.

Aprendizado de Máquina

Habilidade de aprender sem ser explicitamente programado. Aprende com a experiência

Deep Learning

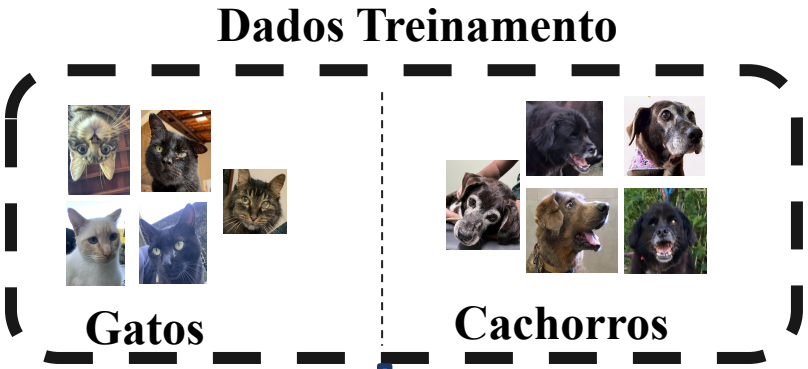
Aprendizado baseado em Redes Neurais Profundas



IA Generativa

1950 ... 1980 ... 2010 2012 ... 2022

Aprendizado de Máquina

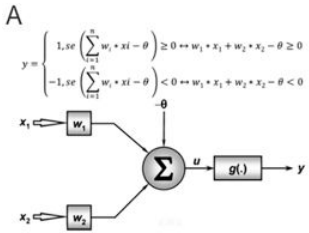


Algoritmo de Aprendizado

Modelo

Gato (conf. 0.96)

Algoritmo de Aprendizado



Os sistemas **de IAGen** são treinados em enormes conjuntos de dados de texto, imagens e outras mídias para produzir conteúdo semelhante, mas sintético.

- Geração de texto, imagens, vídeos, códigos de programação, dados, tradução de idiomas

IA Ética

O que é - por quem

IA Ética

Diversas recomendações e regulações têm sido propostas*
(1) Justiça, diversidade e não discriminação,
(2) Transparência e explicabilidade,
(3) Robustez técnica e segurança,
(4) Privacidade e proteção de dados, e
(5) Responsabilidade e prestação de contas

Princípios para que os sistemas de IA sejam confiáveis, desenvolvidos e utilizados para o bem da humanidade e do planeta e para preservar os valores por meio da proteção, promoção e respeito aos direitos humanos fundamentais, a liberdade e a igualdade.

*Por Ferro M., baseando-se nos pontos comuns encontrados nos documentos (UNESCO, 2021), regulamentação da União Europeia e OCDE (https://oecd.ai/en/classification)

IA Ética

Diversas recomendações e regulações têm sido propostas*
(1) Justiça, diversidade e não discriminação,
(2) Transparência e explicabilidade,
(3) Robustez técnica e segurança,
(4) Privacidade e proteção de dados, e
(5) Responsabilidade e prestação de contas

Princípios para que os sistemas de IA sejam confiáveis, desenvolvidos e utilizados para o bem da humanidade e do planeta e para preservar os valores por meio da proteção, promoção e respeito aos direitos humanos fundamentais, a liberdade e a igualdade.

Crescimento inclusivo
Desenvolvimento sustentável

Futuro do Trabalho

Impactos IA

Exemplos de possíveis impactos

Impactos Negativos IA

- A IA pode impulsionar preconceitos de gênero, raça, idade.

- Alguns sistemas de IA reforçam estereótipos de gênero sobre o trabalho de cuidado e assistência. Por exemplo, a voz feminina de assistentes pessoais virtuais (*Virtual Personal Assistants* – VPAs), como Alexa e Siri, pode reforçar o estereótipo de que as mulheres devem cuidar, se ocupar e assistir ao lar e às necessidades das pessoas da casa (UNESCO; BIF; OCDE, 2023).

- A IA também já foi acusada de impulsionar o ódio às minorias e influenciar os resultados de eleições (Cavaliere; Romeo, 2022), explorar fraquezas psicológicas e orientar decisões (Sartori; Theodorou, 2022), causando problemas como a intensa polarização social e ameaças aos princípios democráticos e aos direitos humanos.

Impactos Negativos IA

Como os sistemas de IA são projetados por seres humanos, é possível que eles injetem seu viés neles, mesmo de maneira não intencional.

Uma maneira predominante de injetar viés pode estar na coleta e seleção de dados de treinamento. Se os dados de treinamento não forem inclusivos, representativos e equilibrados o suficiente, o sistema poderá aprender a **tomar decisões injustas, principalmente em grupos marginalizados, subrepresentados e minorias**.

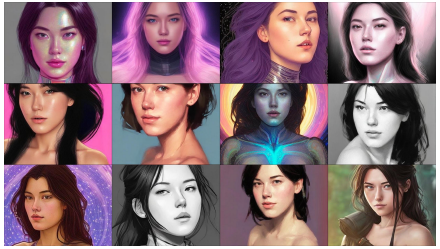
Justiça, Diversidade e Não discriminação

São diversas as situações de discriminação de raça e de vieses e discriminação aos grupos minoritários ou culturas, principalmente com o uso de algoritmos de reconhecimento facial ou manipulação de imagens.

Aplicativo que “desnudava” mulheres mostrando como as *deepfakes* prejudicam os mais vulneráveis (REVIEW, 2022);

Fonte: (REVIEW, 2022);

O algoritmo GAN gerava apenas imagens do corpo feminino, mesmo quando recebia a foto de um homem, pois foi treinado apenas com fotos de mulheres nuas.



Fonte: (Heikkilä, 2022).

Aplicativos que transformam fotos em caricaturas onde os avatares das mulheres, especialmente orientais, são “pornificadas”, enquanto o dos homens são astronautas, exploradores e inventores (Heikkilä, 2022).

Robustez e Proteção de Dados

IA erra e acusa médico de 100 assédios: 'Quase acabou com uma carreira' - Bing, buscador da Microsoft (Uol Tilt, 22/06/23).

A ViaQuatro, empresa que tem a concessão da linha 4-amarela do metrô de São Paulo, foi processada pelo Instituto Brasileiro de Defesa do Consumidor por usar câmeras que coletavam dados referentes às "emoções" dos passageiros, e que seriam usados pela companhia, sem o consentimento dos passageiros (Ferro; Teixeira, 2023).

Ética em PLN

- Monopólio linguístico (inglês) - homogeneização
- Treinamento de grandes modelos de língua tendem a ser compostos por uma quantidade massiva de dados linguísticos coletados online
- Dados que refletem os valores e as normas do Norte Global
- Podem ser não representativos e não levarem em conta a diversidade cultural e linguística existentes.
- Propriedade intelectual e direitos autorais. Transparência, proteção de dados e consentimento são respeitados no processo de coleta?

Regulação IA

Breve panorama regulatório

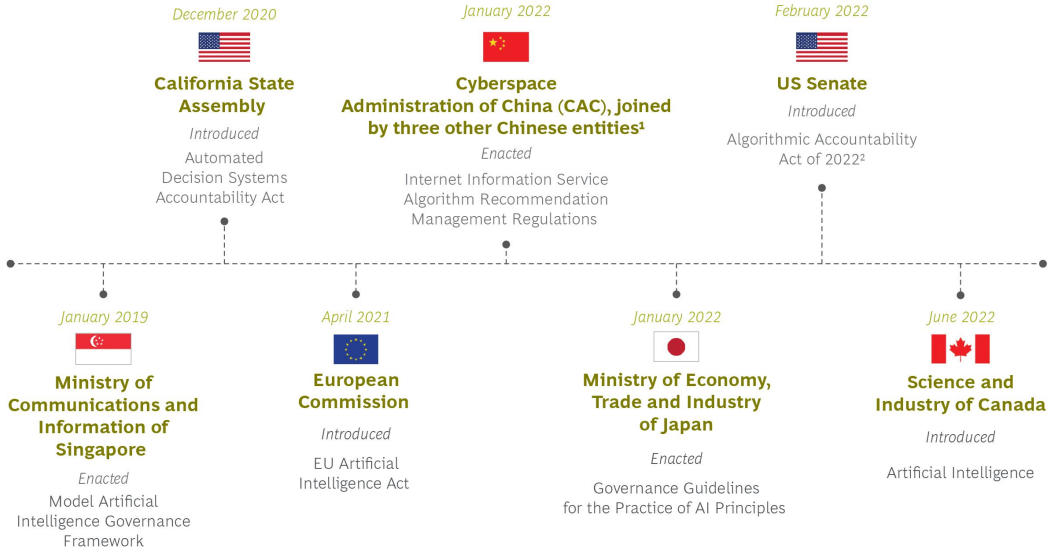
Regulações

Sociedade civil, governo, organizações intergovernamentais, multistakeholder, setor privado

Regulações – Estratégias Nacionais: EU AI Act, China, Canadá, Brasil (EBIA/PL 2338-2023 - 5051/2019, 872/2021 e 21/2020)
Principles - Guidelines –

Observatórios: OCDE, UNESCO, European Union, IBM, Declaração de Montreal, Beijing AI Principles

Standards: IEEE, ISO/IEC 42001, Microsoft



Source: BCG analysis.

Note: Month and year refer to the date of introduction of the law.

[1]Ministry of Industry and Information Technology, Ministry of Public Security, and State Administration for Market Regulation.

[2]Update of 2019 version.

Considerações Finais

Considerações Finais

- A tecnologia, a IA, está gerando novos desafios para a reflexão ética e, consequentemente, para as decisões e ações morais;
- Existem muitos aspectos positivos com a IA, inclusão, solução de problemas complexos da sociedade;
- Impactos sobre a liberdade, diversidade, direitos fundamentais e meio ambiente;
- Ambientes de desenvolvimento e de debate inclusivos e plurais;
- Letramentos em IA, ética e proteção de dados pessoais;
- Conhecer a tecnologia para desenvolver o pensamento crítico sobre como utilizar com ética;
- Tornar a sociedade atuante e parte do processo de desenvolvimento de uma sociedade ética e inclusiva.

Obrigada!

https://brasileiraspln.com/livro-pln/