Contribuição da FENAJUD

Expositor:
Sérgio Amadeu da Silveira
UFABC - CNPq

Três temas:

1) soberania digital e de dados;
2) participação dos trabalhadores;
3) discriminação e vieses;

O cenário

Atualmente, os Estados-nação
e seus poderes são desafiadas pelo alcance
global das **infraestruturas computacionais** e
pelas **redes de comunicação** controladas
por **oligopólios digitais**.

IA realmente existente
IA generativa

Sistemas automatizados baseados em estatística, probabilidade e predição que extraem padrões de dados a partir do seu processamento em infraestruturas de alto poder computacional

A linha de montagem do aprendizado de máquina:

Dados → Algoritmo → Modelo

Alto Poder Computacional

Os dados são o insumo fundamental da IA generativa e da IA realmente existente.


LLM
LARGE LANGUAGE MODEL

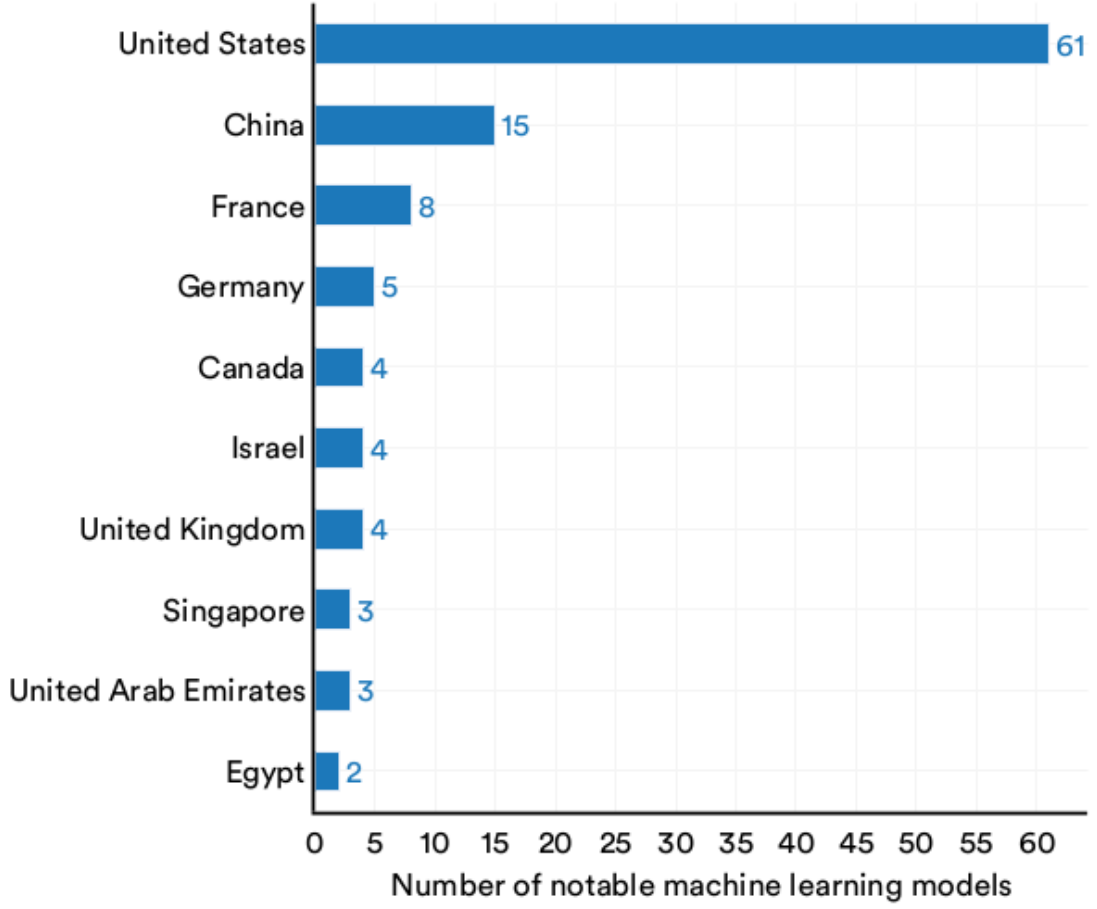
Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

Os oligopólios digitais tentam controlar
os dados e as infraestruturas de treinamento
e processamento dos modelos de IA.

Os provedores de nuvem
são infraestruturas geopolíticas com gigantesca
capacidade de mediação e interferência nas
dinâmicas dos Estados tecnodependentes.

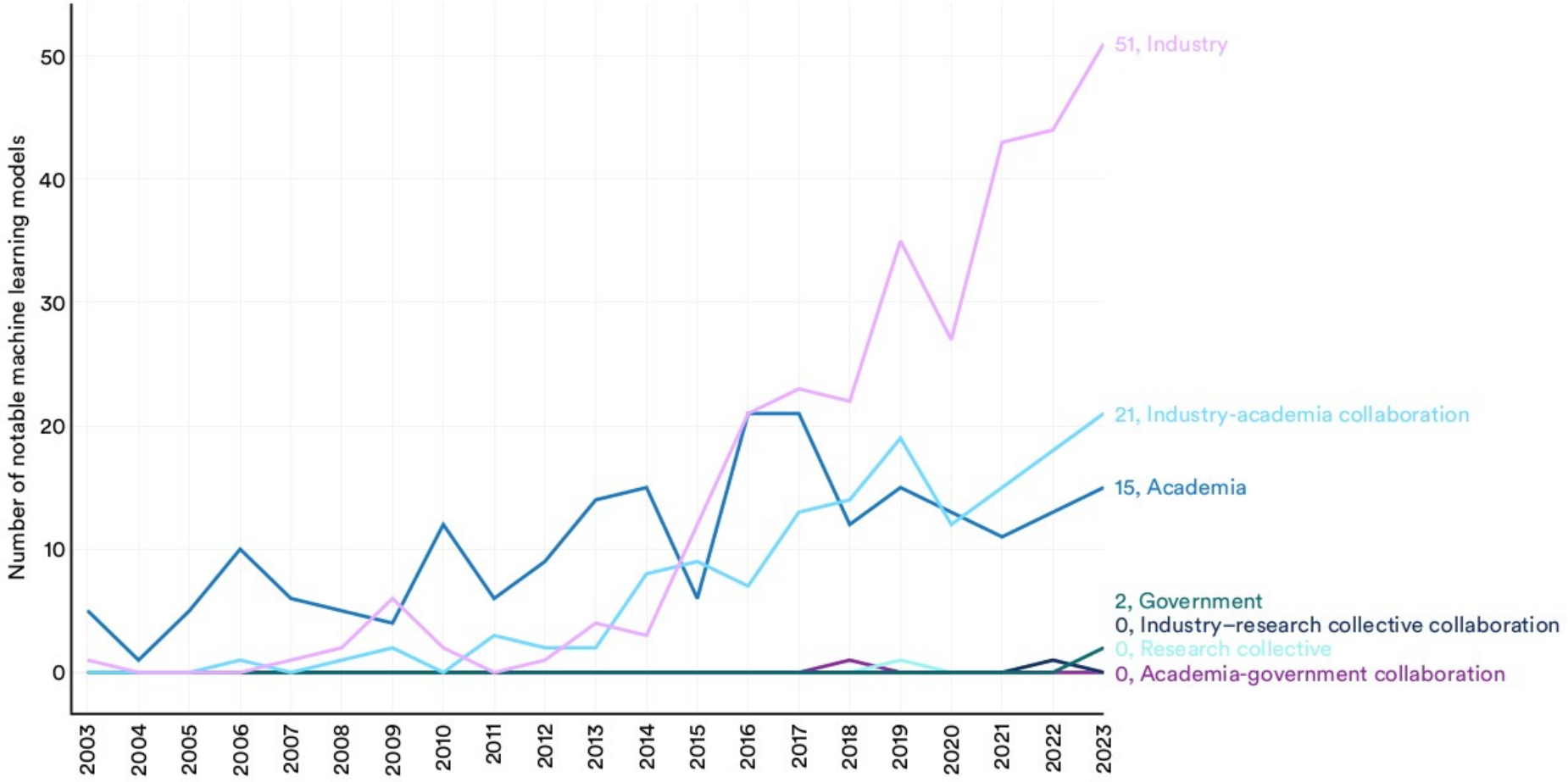
Number of notable machine learning models by geographic area, 2023

Source: Epoch, 2023 | Chart: 2024 AI Index report



Number of notable machine learning models by sector, 2003–23

Source: Epoch, 2023 | Chart: 2024 AI Index report



The privatization of AI research

The results underscore another growing problem in AI, too: the sheer intensity of resources now required to produce paper-worthy results has made it increasingly challenging for people working in academia to continue contributing to research.

LLaMA: Open and Efficient Foundation Language Models

Hugo Touvron[*], **Thibaut Lavril**[*], **Gautier Izacard**[*], **Xavier Martinet**
Marie-Anne Lachaux, Timothee Lacroix, Baptiste Rozière, Naman Goyal
Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin
Edouard Grave[*], **Guillaume Lample**[*]

Meta AI

Abstract

We introduce LLaMA, a collection of foundation language models ranging from 7B to 65B parameters. We train our models on trillions of tokens, and show that it is possible to train state-of-the-art models using publicly available datasets exclusively, without resorting to proprietary and inaccessible datasets. In particular, LLaMA-13B outperforms GPT-3 (175B) on most benchmarks, and LLaMA-65B is competitive with the best models, Chinchilla-70B and PaLM-540B. We release all our models to the research community[1].

1 Introduction

Large Languages Models (LLMs) trained on massive corpora of texts have shown their ability to perform new tasks from textual instructions or from a few examples (Brown et al., 2020). These few-shot properties first appeared when scaling models to a sufficient size (Kaplan et al., 2020), resulting in a line of work that focuses on further scaling these models (Chowdhery et al., 2022; Rae et al., 2021). These efforts are based on the assumption that more parameters will lead to better performance. However, recent work from Hoffmann et al. (2022) shows that, for a given compute budget, the best performances are not achieved by the largest models, but by smaller models trained on more data.

The objective of the scaling laws from Hoffmann et al. (2022) is to determine how to best scale the dataset and model sizes for a particular *training* compute budget. However, this objective disregards the *inference* budget, which becomes critical when serving a language model at scale. In this context, given a target level of performance, the preferred model is not the fastest to train but the fastest at inference, and although it may be cheaper to train a large model to reach a certain level of performance, a smaller one trained longer will ultimately be cheaper at inference. For instance, although Hoffmann et al. (2022) recommends training a 10B model on 200B tokens, we find that the performance of a 7B model continues to improve even after 1T tokens.

The focus of this work is to train a series of language models that achieve the best possible performance at various inference budgets, by training on more tokens than what is typically used. The resulting models, called *LLaMA*, ranges from 7B to 65B parameters with competitive performance compared to the best existing LLMs. For instance, LLaMA-13B outperforms GPT-3 on most benchmarks, despite being 10× smaller. We believe that this model will help democratize the access and study of LLMs, since it can be run on a single GPU. At the higher-end of the scale, our 65B-parameter model is also competitive with the best large language models such as Chinchilla or PaLM-540B.

Unlike Chinchilla, PaLM, or GPT-3, we only use publicly available data, making our work compatible with open-sourcing, while most existing models rely on data which is either not publicly available or undocumented (e.g. "Books – 2TB" or "Social media conversations"). There exist some exceptions, notably OPT (Zhang et al., 2022), GPT-NeoX (Black et al., 2022), BLOOM (Scao et al., 2022) and GLM (Zeng et al., 2022), but none that are competitive with PaLM-62B or Chinchilla.

In the rest of this paper, we present an overview of the modifications we made to the transformer architecture (Vaswani et al., 2017), as well as our training method. We then report the performance of our models and compare with others LLMs on a set of standard benchmarks. Finally, we expose some of the biases and toxicity encoded in our models, using some of the most recent benchmarks from the responsible AI community.

[*] Equal contribution. Correspondence: {htouvron, thibautlav,gizacard,egrave,glample}@meta.com
[1] https://github.com/facebookresearch/llama

Language Models are Few-Shot Learners

Tom B. Brown[*] **Benjamin Mann**[*] **Nick Ryder**[*] **Melanie Subbiah**[*]
Jared Kaplan[†] **Prafulla Dhariwal** **Arvind Neelakantan** **Pranav Shyam** **Girish Sastry**
Amanda Askell **Sandhini Agarwal** **Ariel Herbert-Voss** **Gretchen Krueger** **Tom Henighan**
Rewon Child **Aditya Ramesh** **Daniel M. Ziegler** **Jeffrey Wu** **Clemens Winter**
Christopher Hesse **Mark Chen** **Eric Sigler** **Mateusz Litwin** **Scott Gray**
Benjamin Chess **Jack Clark** **Christopher Berner**
Sam McCandlish **Alec Radford** **Ilya Sutskever** **Dario Amodei**

OpenAI

Abstract

Recent work has demonstrated substantial gains on many NLP tasks and benchmarks by pre-training on a large corpus of text followed by fine-tuning on a specific task. While typically task-agnostic in architecture, this method still requires task-specific fine-tuning datasets of thousands or tens of thousands of examples. By contrast, humans can generally perform a new language task from only a few examples or from simple instructions – something which current NLP systems still largely struggle to do. Here we show that scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even reaching competitiveness with prior state-of-the-art fine-tuning approaches. Specifically, we train GPT-3, an autoregressive language model with 175 billion parameters, 10x more than any previous non-sparse language model, and test its performance in the few-shot setting. For all tasks, GPT-3 is applied without any gradient updates or fine-tuning, with tasks and few-shot demonstrations specified purely via text interaction with the model. GPT-3 achieves strong performance on many NLP datasets, including translation, question-answering, and cloze tasks, as well as several tasks that require on-the-fly reasoning or domain adaptation, such as unscrambling words, using a novel word in a sentence, or performing 3-digit arithmetic. At the same time, we also identify some datasets where GPT-3's few-shot learning still struggles, as well as some datasets where GPT-3 faces methodological issues related to training on large web corpora. Finally, we find that GPT-3 can generate samples of news articles which human evaluators have difficulty distinguishing from articles written by humans. We discuss broader societal impacts of this finding and of GPT-3 in general.

[*] Equal contribution
[†] Johns Hopkins University, OpenAI

Author contributions listed at end of paper.

Em **julho de 2020**, a OpenAI revelou o **GPT-3**,
o maior modelo de linguagem então conhecido.

O GPT-3 possui **175 bilhões de parâmetros** e foi treinado
em **570 gigabytes de texto**. Para efeito de comparação,
seu antecessor, **GPT-2, era 100 vezes menor**,
com **1,5 bilhão de parâmetros**.

Fonte: THE AI INDEX REPORT 2024

O **GPT-4** da OpenAI usou cerca de **US$ 78 milhões** em computação para treinar, enquanto o **Gemini Ultra** do Google custou **US$ 191 milhões** em computação.

Fonte: THE AI INDEX REPORT 2024

O treinamento de um modelo simples pode ser obtido com hardware comum , mas os obstáculos para **a implantação de um LLM** (large language models) de **última geração são significativos**.

Os oligopólios digitais tentam controlar
os **dados** e as **infraestruturas de treinamento**
e processamento dos **modelos de IA**.

Os **provedores de nuvem**
são **infraestruturas geopolíticas** com
gigantesca capacidade de **interferência nas dinâmicas** dos **Estados** tecnodependentes.

Ao controlar os dados e as tecnologias indispensáveis, as **Big Techs** vão **controlando** as **dinâmicas das organizações**.

A **concentração no mercado de armazenagem de dados** e dos chamados **serviços de nuvem**, com apenas cinco empresas detendo **81,2%** do mercado mundial de **Infraestrutura como Serviço** (IaaS) em **2021.**

Estas empresas são **Amazon (38,9%)**, **Microsoft (21,1%)**, Alibaba (9,5%), Google (7,1%), e Huawei (4,6%).

Em 2022, a **Amazon**
aumentou sua participação para **40%** e a
Microsoft Azure para **21,5%**, totalizando
juntas **61,5% do mercado de IaaS**

Google Play

Pesquisar

Apps

Categorias

Pessoal

Em alta

Lançamentos

Meus apps

Comprar

Jogos

Família

Escolha dos editores

Conta

Formas de pagamento

Minhas assinaturas

Resgatar

Comprar vale-presente

Minha lista de desejos

Minha atividade Play

Guia para a família

SouGov.br

Governo do Brasil Finanças

2.351

Este app está disponível para seu dispositivo

Adicionar à lista de desejos

Instalar

Semelhantes

Veja mais

gov.br
Governo do Brasil
O gov.br é um meio de acesso do usuário aos serviços públicos digitais

Consumidor.gov
Governo do Brasil
O Consumidor.gov.br é um serviço público para solução de conflitos de

Bradesco Seguro
Bradesco Seguros S.A
Bradesco Seguros: Seus produtos e planos em um único app

Transferência internacional de dados

SouGov utiliza a solução de central de ajuda (chat), denominada SerproBot, que utiliza tecnologia da empresa IBM - International Business Machines. Nesse contexto, o usuário fica ciente de que os dados digitados no chat poderão ser transferidos internacionalmente e ficam armazenados na infraestrutura da empresa por um período de 30 (trinta) dias. Após este período os dados são excluídos em definitivo. Tal armazenamento tem o objetivo de prover o aprendizado de máquina da ferramenta de chat denominada "Watson", onde as interações dos usuários no chat são utilizadas para "aprendizado" pelo computador que envia as respostas automáticas quando o usuário está sendo atendido por meio do chat do serviço SouGov.

Ao concordar com Termo de Uso e Política de Privacidade do aplicativo SouGov o usuário estará consentindo com a transferência internacional das informações digitadas no chat do aplicativo SouGov.

Como o chat está programado apenas para responder questões sobre a navegação e funcionalidades do aplicativo, os usuários não são orientados e n
há necessidade de digitar qualquer informação de caráter pessoal na interação por meio do chat.

País: Estados Unidos da América, Organização: IBM - International Business Machines.

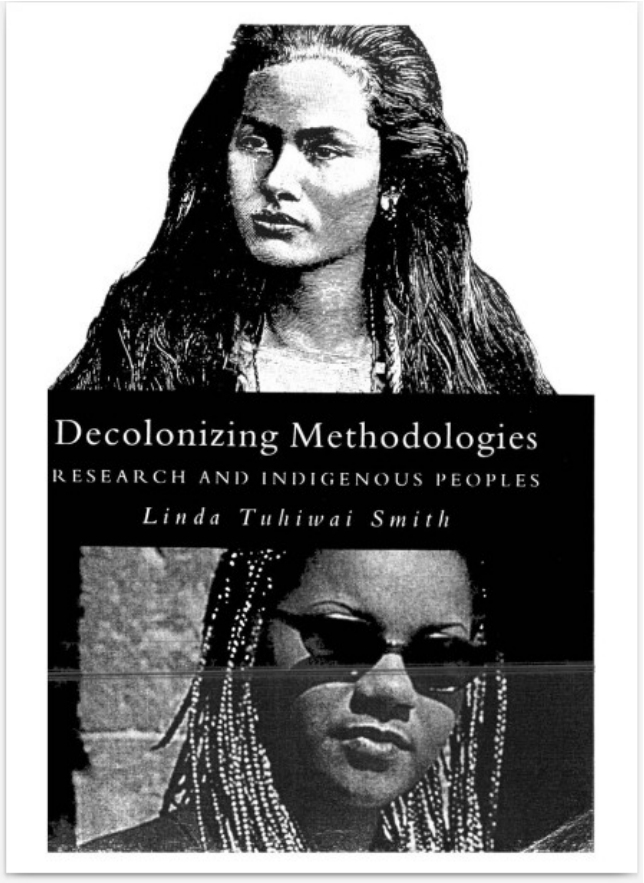
"… Tal armazenamento tem o objetivo de **prover o aprendizado de máquina da ferramenta de chat denominada "Watson"**, onde as **interações** dos usuários no chat são utilizadas para "aprendizado" pelo computador que envia as respostas automáticas quando o usuário está sendo atendido por meio do chat do serviço SouGov."

Termos de Uso do SouGov

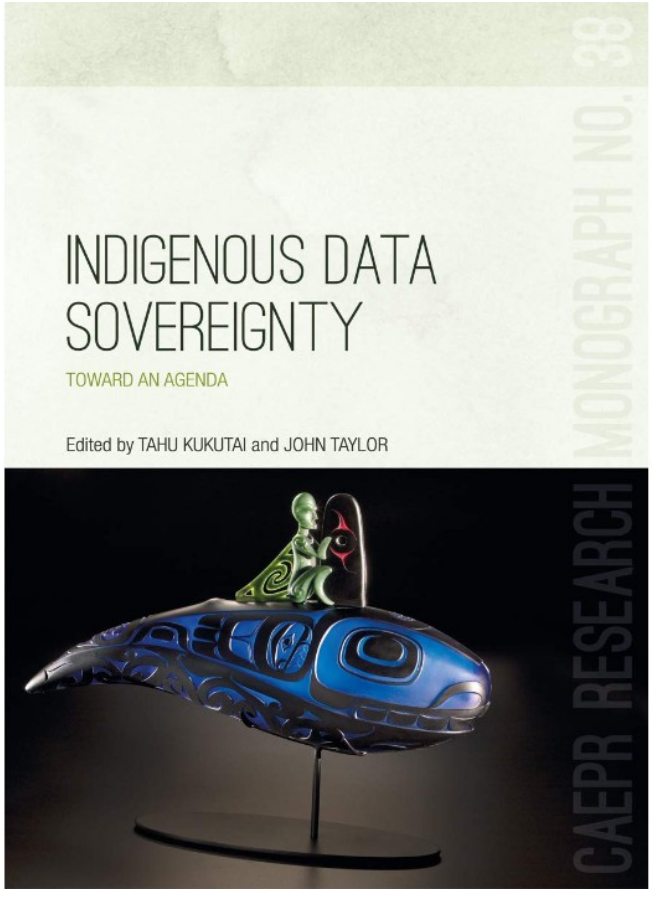

POWER7
POWER7
POWER7
WATSON
POWER7
POWER7
POWER7

Soberania digital é a capacidade de
uma sociedade e seu Estado de definir, governar e controlar
as tecnologias indispensáveis à sua autodeterminação,
à proteção dos seus direitos, à sua inventividade,
tecnodiversidade e desenvolvimento.

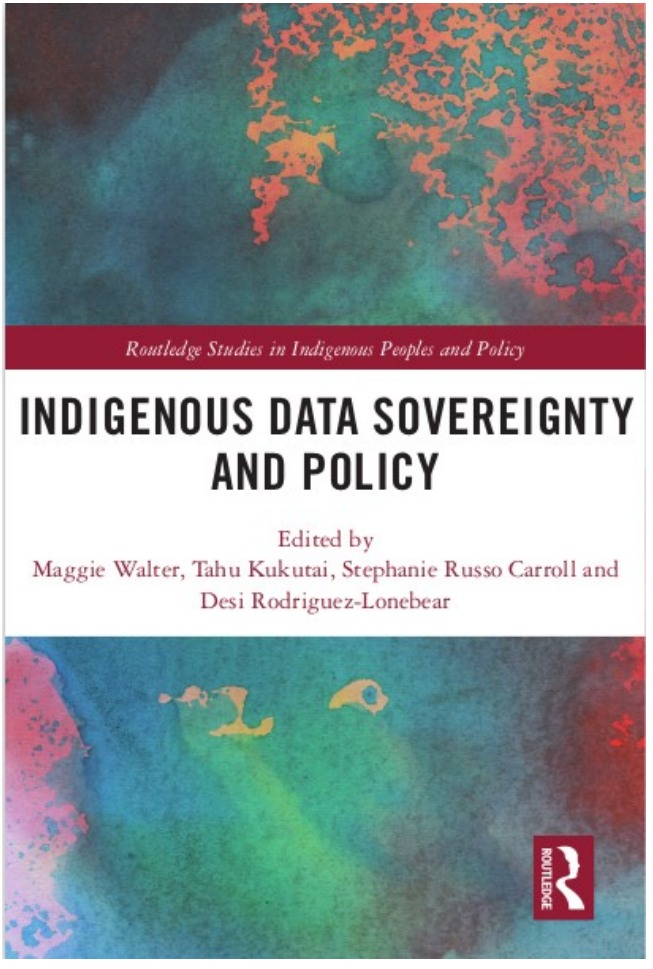
O termo **"soberania alimentar"** foi cunhado pela primeira vez em **1996** por membros da **Via Campesina** , uma organização internacional de agricultores, e posteriormente adotado por várias organizações internacionais, incluindo o **Banco Mundial** e as **Nações Unidas**.



1999



2016



2021

Soberania de dados é o poder de uma sociedade **decidir quais dados podem ser criados**, como devem ser armazenados, processados, reutilizados, analisados, e com quais finalidades, em qualquer tempo.

Oracle Launches EU Sovereign Cloud to Reinforce Data Sovereignty and Privacy

Microsoft launches 'sovereign' cloud for governments

By **Supantha Mukherjee**

July 19, 2022 12:05 PM GMT-3 · Updated a year ago

FENAJUD propõe que o CNJ defina:

1- Os Tribunais e órgãos do Judiciário brasileiro **não possam armazenar dados fora de nossa Jurisdição.**

2- Que os **dados estratégicos e sensíveis do Judiciário** brasileiro **não possam ser utilizados para treinar modelos** de IA, generativa ou não, para empresas sob o controle de capitais estrangeiros.

FENAJUD propõe que o CNJ defina:

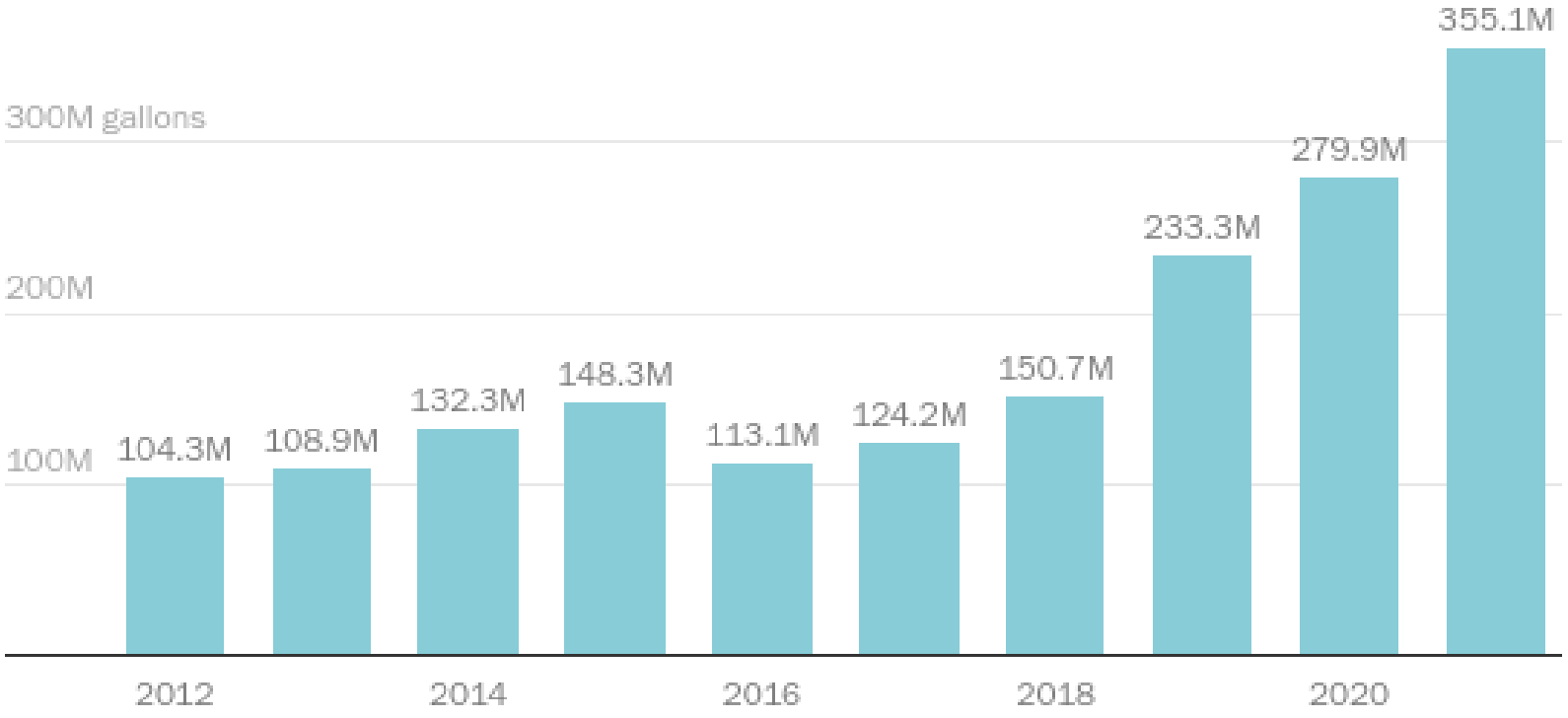
3- Que dados considerados sensíveis ou estratégicos não possam estar sob o controle ou guarda de empresas transnacionais que tenham notórios interesses econômicos, políticos ou geopolíticos no país.

4- Que a soberania digital e de dados seja inserida como princípio e fundamento da resolução do CNJ.

Que o Poder Judiciário implemente a Justiça Ambiental

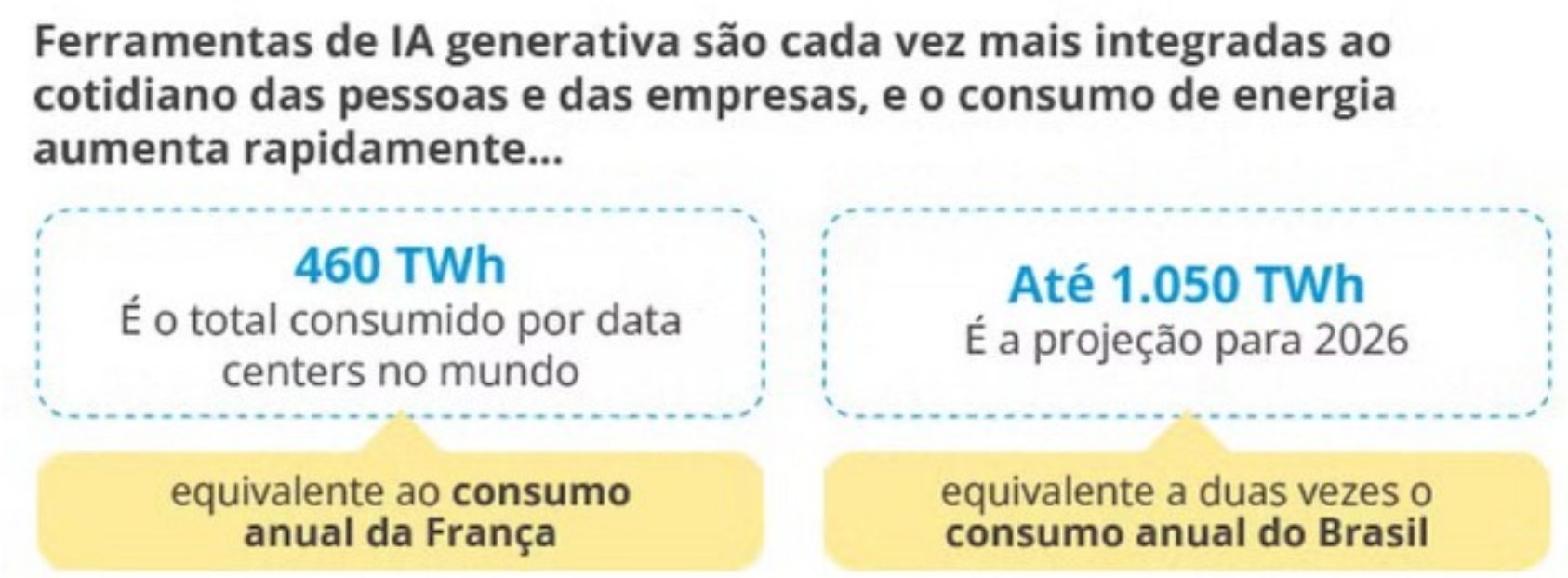
Google's annual water use in The Dalles, in gallons

Google's data centers in The Dalles use nearly three times more water than they did five years ago and now account for more than a quarter of all the city's water use.



Source: City of The Dalles • Get the data

OREGONLIVE
The Oregonian



Fonte: Inteligência artificial exigirá energia de ‘dois Brasis’ até 2026.
Link: https://visaosocioambiental.com.br/inteligencia-artificial-exigira-energia-de-dois-brasis-ate-2026-veja-por-que-a-tecnologia-demanda-tanto/

Para manter a temperatura adequada para seus milhares de servidores, os **Data Centers** já são responsáveis por **2,5** a **3,7** % da **emissões** globais de **gases do efeito estufa**, superando a indústria de aviação.

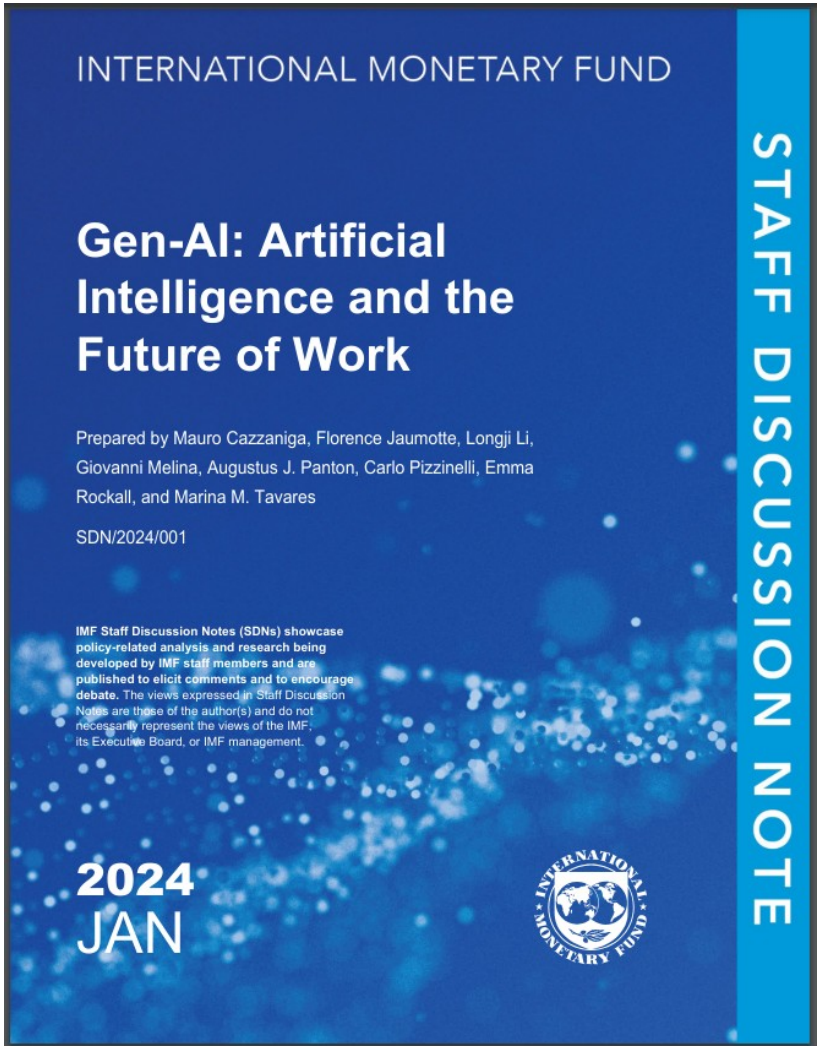
Os data centers são responsáveis por 2,5% a 3,7% das emissões globais de GEE.

insustentável

A FENAJUD defende que o **Poder Judiciário** utilize **infraestruturas soberanas de armazenamento,** processamento e análise de dados, bem como, treinem seus modelos de IA utilizando **tecnologias abertas** e **infraestruturas distribuídas** em nosso território que **priorizem o baixo impacto ambiental**.

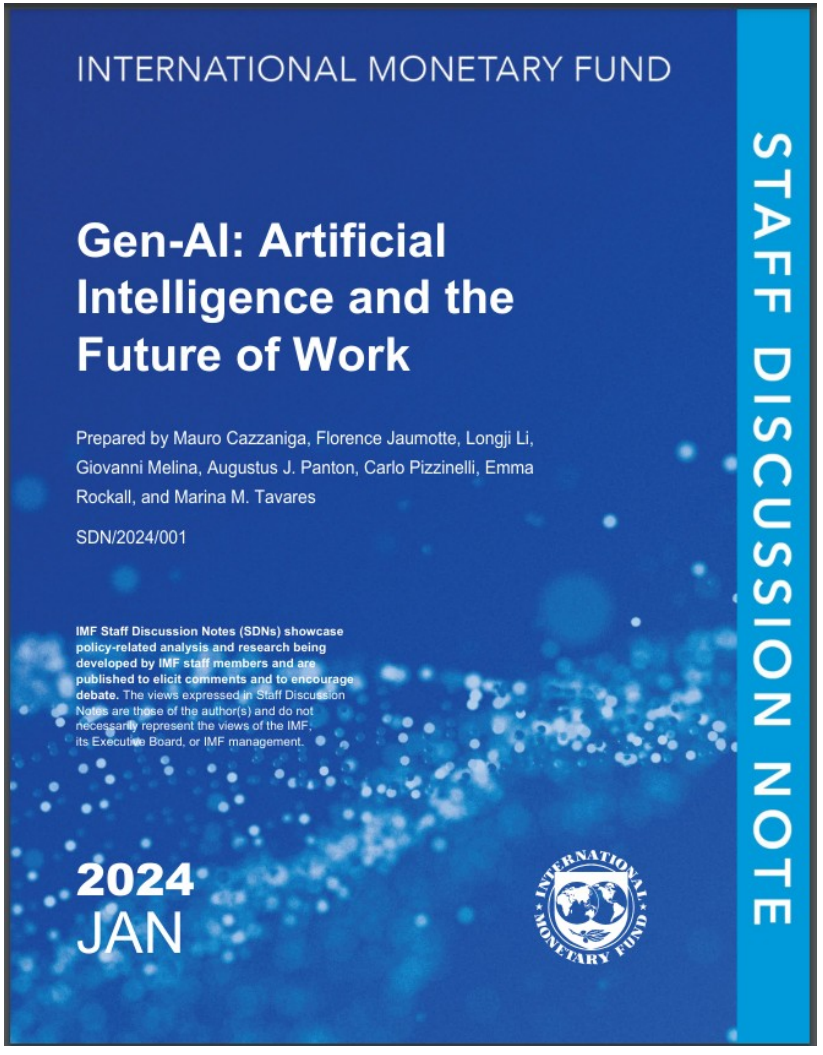
FENAJUD

Para a FENAJUD, o CNJ deve inserir como regra de desenvolvimento e uso de IA no Judiciário a **avaliação do impacto ambiental**, bem como, a necessidade de redução do **consumo de energia** e da **pegada de carbono da IA implementada**, calculada no mínimo a cada três anos.



"Nas economias avançadas, cerca de **60 por cento dos empregos estão expostos à IA**, devido à prevalência de empregos orientados para tarefas cognitivas."

“A exposição geral é de **40% nas economias de mercado emergentes** e **26%** nos **países de baixa renda**.”



OCTOBER 30, 2023

Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

By the authority vested in me as President by the Constitution and the laws of the United States of America, it is hereby ordered as follows:

Section 1. Purpose. Artificial intelligence (AI) holds extraordinary potential for both promise and peril. Responsible AI use has the potential to help solve urgent challenges while making our world more prosperous, productive, innovative, and secure. At the same time, irresponsible use could exacerbate societal harms such as fraud, discrimination, bias, and disinformation; displace and disempower workers; stifle competition; and pose risks to national security. Harnessing AI for good and realizing its myriad benefits requires mitigating its substantial risks. This endeavor demands a society-wide effort that includes government, the private sector, academia, and civil society.

“Os próximos passos críticos no desenvolvimento da **IA devem ser construídos com base nas visões dos trabalhadores, sindicatos, educadores e empregadores** para apoiar usos responsáveis da IA que melhorem a vida dos trabalhadores, aumentem positivamente o trabalho humano e ajudem todas as pessoas a aproveitar com segurança os ganhos e oportunidades da inovação tecnológica.”

Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, 2023

a FENAJUD propõe:

1- Todo Tribunal deve constituir uma **comissão para a implementação da IA e de sistemas automatizados** que tenha pelo menos **um terço de representantes dos trabalhadores** em sua composição.

2- Todo projeto de utilização, desenvolvimento, implementação de IA deve ter a **aprovação da comissão de implementação de IA**.

a FENAJUD propõe:

3- Os Tribunais que forem utilizar, desenvolver ou adquirir a IA devem apresentar um **relatório de impactos nas atividades e postos de trabalho**.

Esse relatório deve conter medidas para mitigar os efeitos nocivos nos empregos e no bem-estar dos funcionários.

a FENAJUD propõe:

4- Os sistemas e IA desenvolvidos, adquiridos, implementados e utilizados devem ter pelo menos **um responsável humano por acompanhar suas atividades, providenciar a correção de seus erros e viéses**.

Os nomes dos responsáveis devem ser públicos e seu contato deve ser facilitado para o público interno e externo ao Tribunal.

Bernard Parker, left, was rated

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

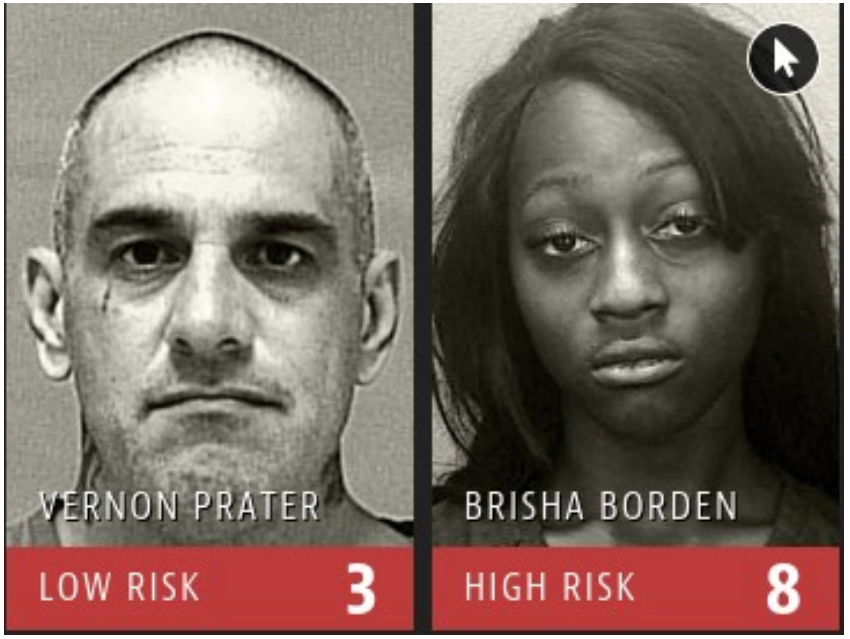
Just as the 18-year-old girls were realizing they were too big for the tiny conveyances — which belonged to a 6-year-old boy — a woman came running after them saying, "That's my kid's stuff." Borden and her friend immediately dropped the bike and scooter and walked away.

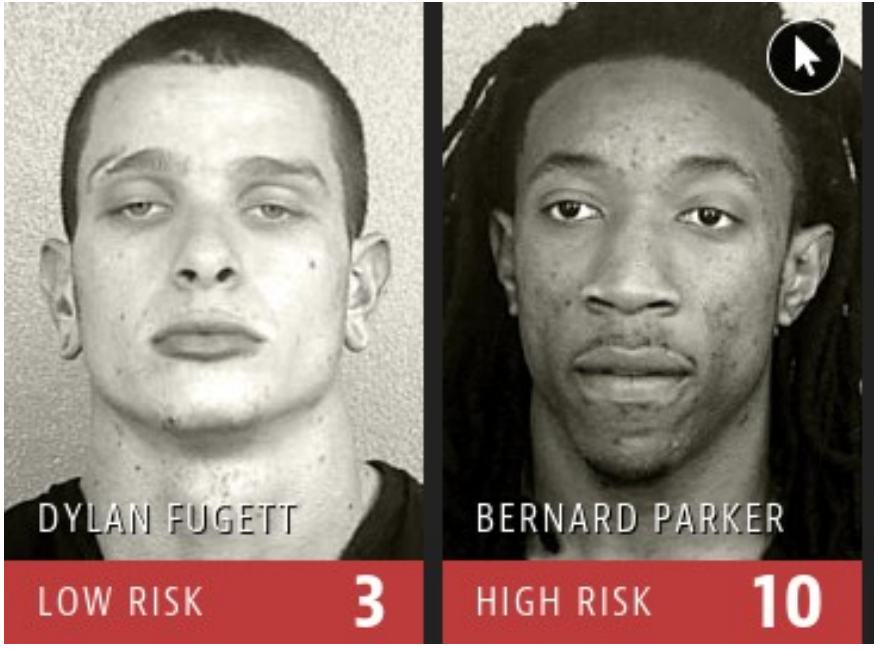
But it was too late — a neighbor who witnessed the heist had already called the police. Borden and her friend were arrested and charged with burglary and petty theft for the items, which were valued at a total of $80.

Vieses e Discriminações

- Três tipos de vieses:
histórico, da base de dados e algorítmico.

Exemplos: exclusão de mulheres pela IA da Amazon e discriminação racial pelo Compas.


VERNON PRATER
LOW RISK 3
BRISHA BORDEN
HIGH RISK 8


DYLAN FUGETT
LOW RISK 3
BERNARD PARKER
HIGH RISK 10

1- O Judiciário **não deve utilizar sistemas algorítmicos** ou automatizados, chamados de inteligentes ou não, entre eles, a IA generativa, **para a tomada de decisões judiciais**, devido ao elevado risco para os direitos e garantias individuais e coletivas.

2- Todo projeto e desenho de implementação da IA no Judiciário deve definir previamente os possíveis **riscos de vieses algorítmicos e de dados** em um **relatório específico que contenha medidas de prevenção, mitigação, correção e reparação** em todas as fases do ciclo de desenvolvimento, implementação e uso da IA.

3- **Todos os sistemas algorítmicos**, incluindo a IA generativa e os modelos de aprendizado de máquina, devem g**arantir a explicabilidade de seus resultados** caso afetem os direitos e garantias individuais e coletivas.

Obrigado.

sergio.amadeu@ufabc.edu.br

SAMADEU