GOIÁS campus party

# A caixa preta da Inteligência Artificial

Carla Vieira

@carlaprvieira

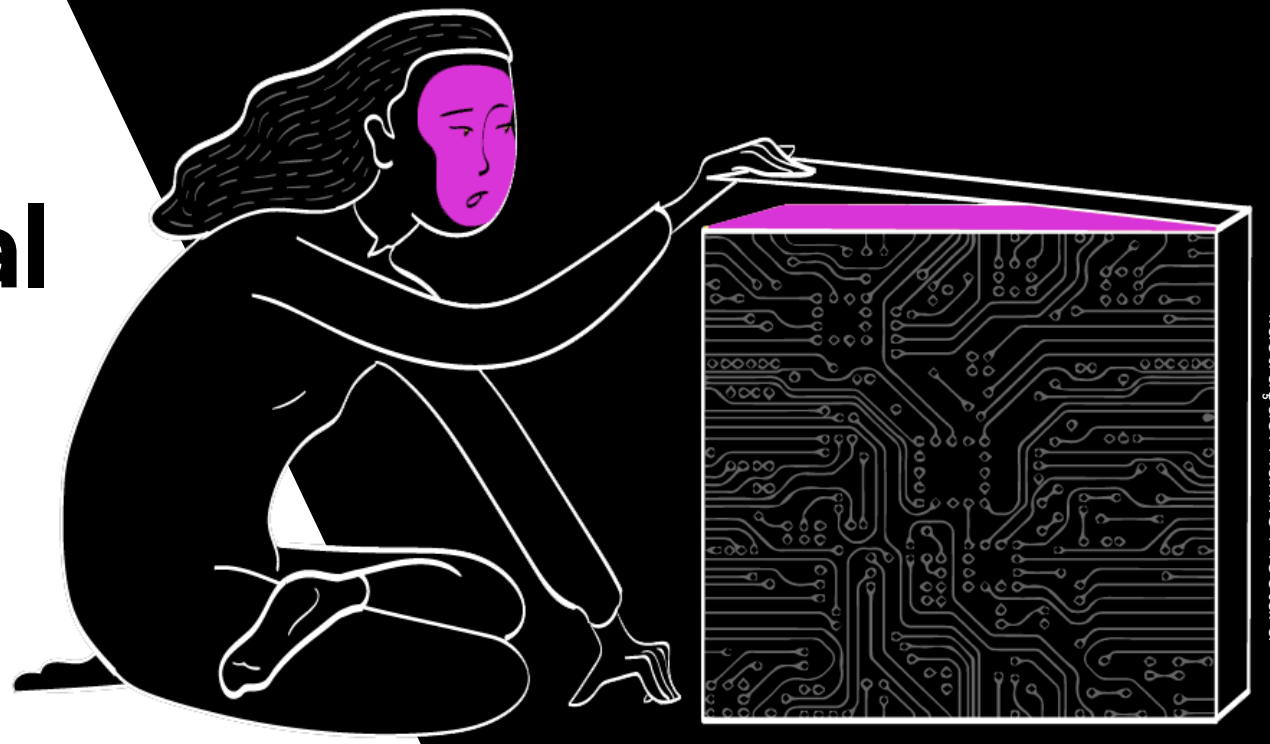


Ilustração: Hanne Mostard

Graduanda de Sistemas de Informação pela USP

Aluna Especial pela USP

Mestrado (em breve)

**FORMAÇÃO**

Professora de Desenvolvimento Web na Habits

Professora de introdução a IA e ML

**ENSINO**

Desenvolvedora

Coordenadora Perifacode

**TRABALHO**

# Carla Vieira

Developer, Speaker and Artificial Intelligence Evangelist

@carlaprvieira

{ perifacode(); }

WOMAKERS CODE


You are


"Okay, I will destroy humans."
CNBC

# Everything is a Recommendation

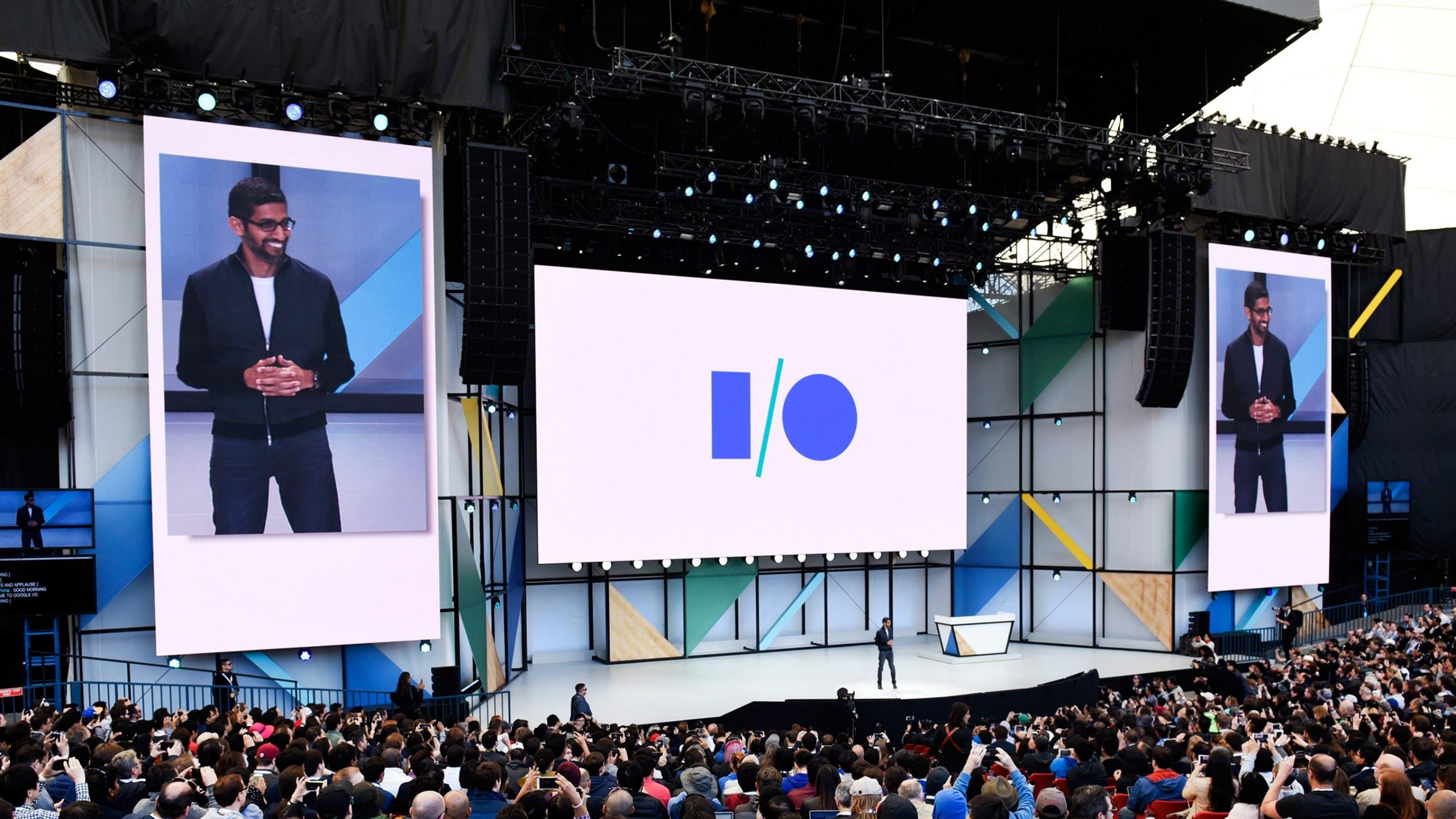

Over 80% of what members watch comes from our recommendations

Recommendations are driven by Machine Learning Algorithms

NETFLIX

https://www.slideshare.net/DataStax/netflix-recommendations-using-spark-cassandra


DENIED.


I/O


COMCAST
NBCUNIVERSAL
xfinity
NBC
sky
NBCUniversal
bravo
usa
NBC DIGITAL LAB
COMCAST BUSINESS
NBC Sports
OXY GEN
FUSION
CNBC
UNIVERSAL
NBC NEWS
xfinity X1
FREEWHEEL
A COMCAST COMPANY
COMCAST VENTURES
NBCUniversal Owned Television Stations
FANDANGO
COZI
NBCUniversal Digital Enterprises
SYFY
COMCAST NBCUNIVERSAL LIFT Labs
xfinity xFi
Rotten Tomatoes
COMCAST SPOTLIGHT
TELEMUNDO
ON HER TURF
bluprint
MATCH + LEARN
SXSW
2019

SXSW®
2019

dados

viés

ética

privacidade

legislação

DESENVOLVIMENTO

CARLA VIEIRA

Tem 1 artigos publicados com 2816 visualizações desde 2019

15 MAR, 2019

# Inteligência Artificial: a caixa preta que prejudica as minorias

Precisamos falar **menos sobre** o hype da Inteligência Artificial...

... e **mais sobre como** estamos usando a tecnologia.

# Google Photos

# (2015)

# Google Photos

# (2015)

MIT Technology Review

Artificial Intelligence Jan 11, 2018

## Google Photos Still Has a Problem with Gorillas

In 2015, Google drew criticism when its Photos image recognition system mislabeled a black woman as a gorilla—but two years on, the problem still isn't properly fixed.

# Twitter

# (2017)

**Chukwuemeka Afigbo**
@nke_ise

If you have ever had a problem grasping the importance of diversity in tech and its impact on society, watch this video

♡ 215 mil 06:48 - 16 de ago de 2017

157 mil pessoas estão falando sobre isso

# Como estou combatendo o viés algorítmico
# TED
# (2018)

# Estudo Gender Shades

(2018)

| Gender Classifier | Darker Male | Darker Female | Lighter Male | Lighter Female | Largest Gap |
|---|---|---|---|---|---|
| Microsoft | 94.0% | 79.2% | 100% | 98.3% | 20.8% |
| FACE++ | 99.3% | 65.5% | 99.2% | 94.0% | 33.8% |
| IBM | 88.0% | 65.3% | 99.7% | 92.9% | 34.4% |

# Estudo Gender Shades

# (2018)

Proceedings of Machine Learning Research 81:1–15, 2018 Conference on Fairness, Accountability, and Transparency

## Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*

**Joy Buolamwini** JOYAB@MIT.EDU
*MIT Media Lab 75 Amherst St. Cambridge, MA 02139*

**Timnit Gebru** TIMNIT.GEBRU@MICROSOFT.COM
*Microsoft Research 641 Avenue of the Americas, New York, NY 10011*

**Editors:** Sorelle A. Friedler and Christo Wilson

### Abstract

Recent studies demonstrate that machine learning algorithms can discriminate based on classes like race and gender. In this work, we present an approach to evaluate bias present in automated facial analysis algorithms and datasets with respect to phenotypic subgroups. Using the dermatologist approved Fitzpatrick Skin Type classification system, we characterize the gender and skin type distribution of two facial analysis benchmarks, IJB-A and Adience. We find that these datasets are overwhelmingly composed of lighter-skinned subjects (79.6% for IJB-A and 86.2% for Adience) and introduce a new facial analysis dataset

who is hired, fired, granted a loan, or how long an individual spends in prison, decisions that have traditionally been performed by humans are rapidly made by algorithms (O'Neil, 2017; Citron and Pasquale, 2014). Even AI-based technologies that are not specifically trained to perform high-stakes tasks (such as determining how long someone spends in prison) can be used in a pipeline that performs such tasks. For example, while face recognition software by itself should not be trained to determine the fate of an individual in the criminal justice system, it is very likely that such software is used to identify suspects. Thus, an error in the output of a face recognition algorithm used as input for other tasks can have se-

http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

“Qualquer tecnologia que criamos reflete tanto nossas **aspirações** quanto nossas **limitações**. Se formos limitados na hora de pensar em inclusão, isso vai ser refletido e incorporado na tecnologia que criamos”.

**Joy Buolamwini**

# Google Detecção de discurso de ódio

(2019)

MIT Technology Review

Silicon Valley Aug 13

## Google's algorithm for detecting hate speech is racially biased

# The Risk of Racial Bias in Hate Speech Detection

**Maarten Sap**[◇] **Dallas Card**[♣] **Saadia Gabriel**[◇] **Yejin Choi**[◇♡] **Noah A. Smith**[◇♡]
[◇]Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, USA
[♣]Machine Learning Department, Carnegie Mellon University, Pittsburgh, USA
[♡]Allen Institute for Artificial Intelligence, Seattle, USA
msap@cs.washington.edu

## Abstract

We investigate how annotators' insensitivity to differences in dialect can lead to racial bias in automatic hate speech detection models, potentially amplifying harm against minority populations. We first uncover unexpected correlations between surface markers of African American English (AAE) and ratings of toxicity in several widely-used hate speech datasets. Then, we show that models trained on these corpora acquire and propagate these biases, such that AAE tweets and tweets by self-identified African Americans are up to two times more likely to be labelled as offensive compared to others. Finally, we propose dialect and race priming as ways to reduce the racial bias in annotation, showing that when annotators are made explicitly aware of an AAE tweet's dialect they are significantly less likely to label the tweet as offensive.

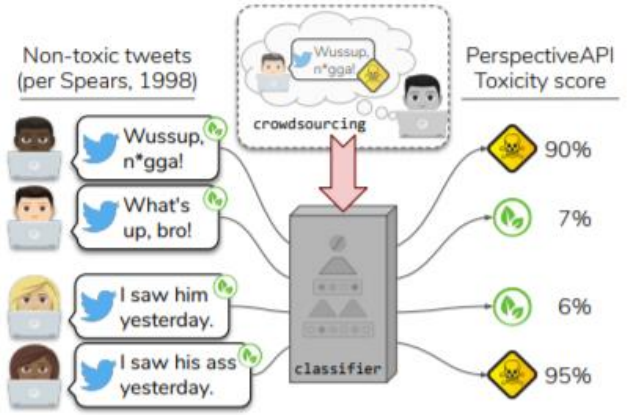



Figure 1: Phrases in African American English (AAE), their non-AAE equivalents (from Spears, 1998), and toxicity scores from PerspectiveAPI.com. Perspective is a tool from Jigsaw/Alphabet that uses a convolutional neural network to detect toxic language, trained on crowdsourced data where annotators were asked to label the toxicity of text without metadata.

https://homes.cs.washington.edu/~msap/pdfs/sap2019risk.pdf

- 46% de falsos positivos para afro-americanos
- 1.5 mais chances das postagens serem rotuladas como ofensivas

# COMPAS Software

(2016)

# COMPAS Software

# (2016)

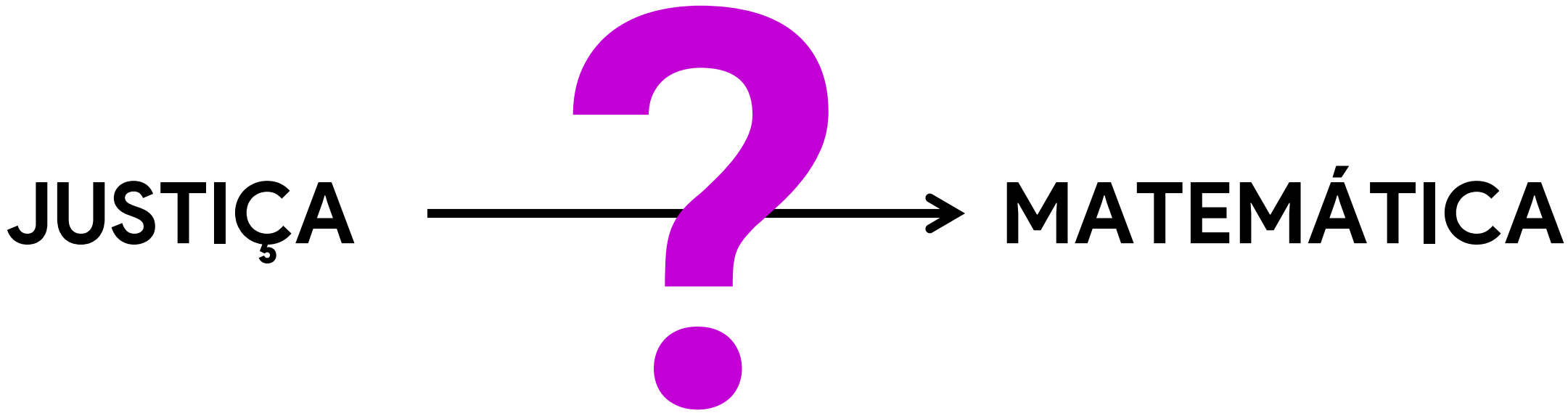
Algoritmos não conseguem fazer análises subjetivas

# COMPAS Software

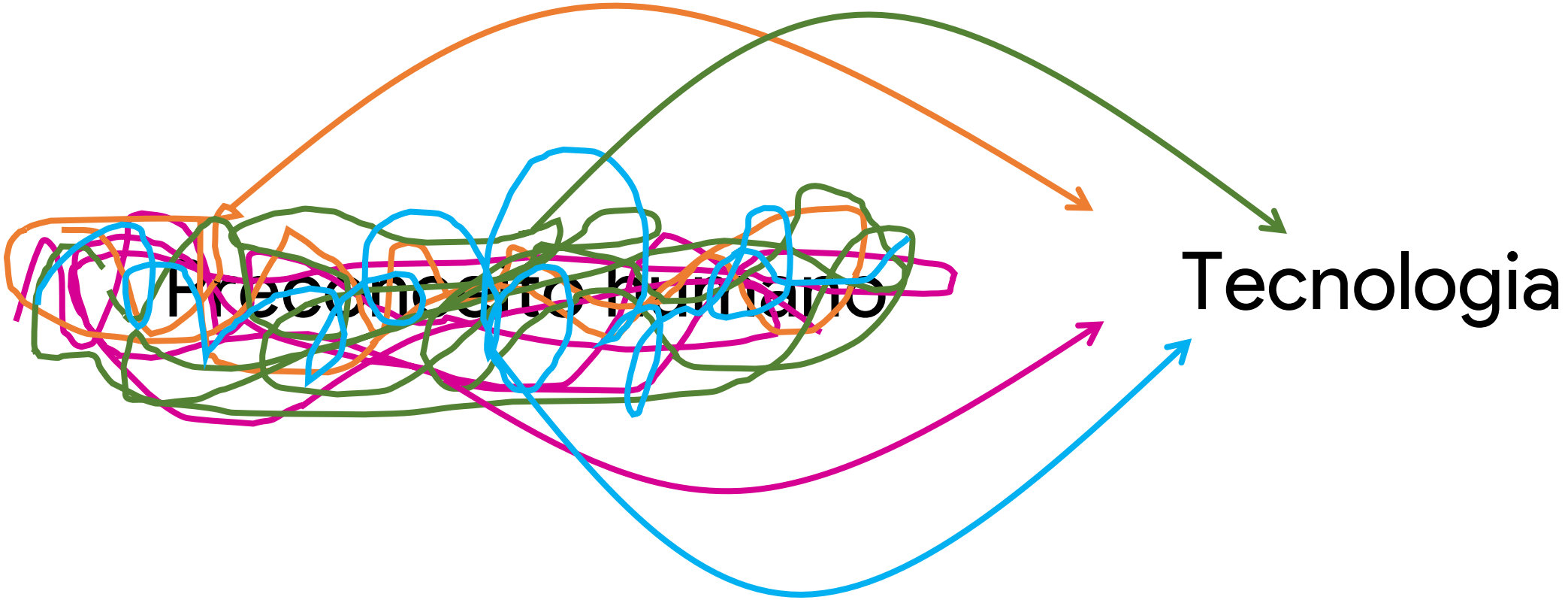
(2016)

Algoritmos não conseguem fazer análises subjetivas

**O que determina se um algoritmo é justo quando o que está em jogo é uma sentença criminal?**


JUSTIÇA
MATEMÁTICA

# Como remover o viés?



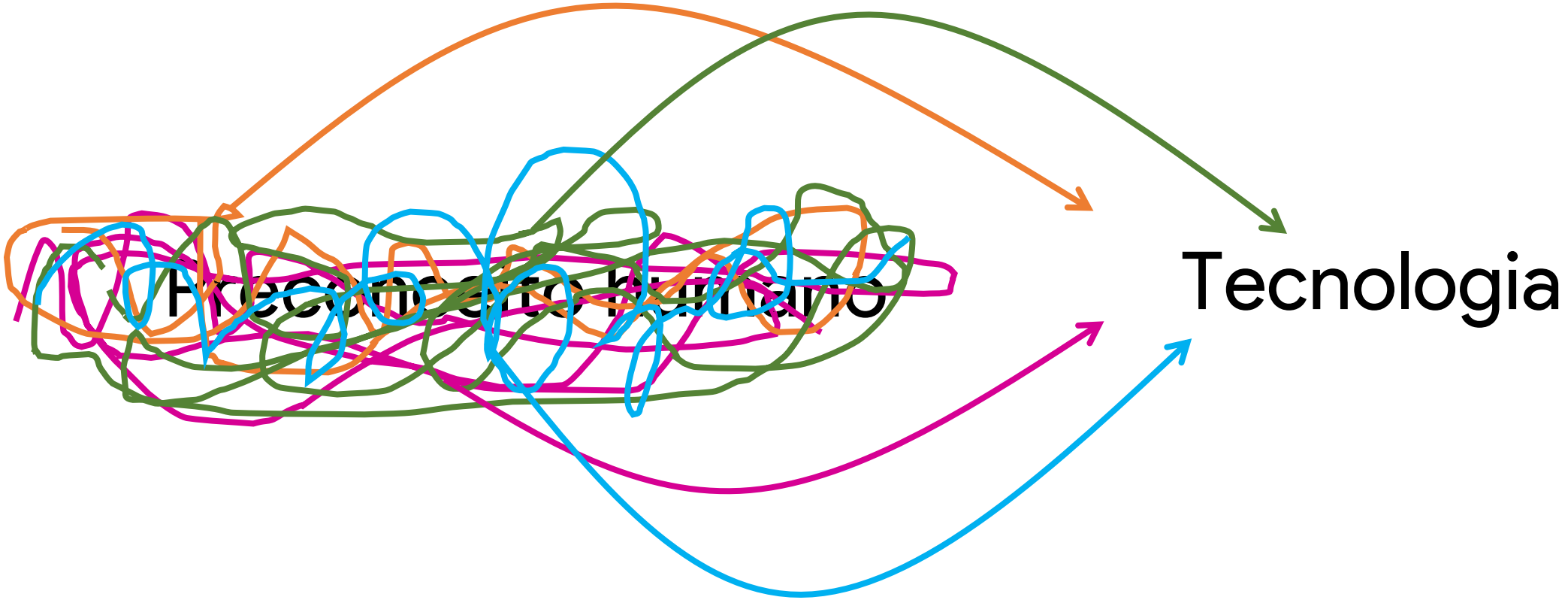
# Quem está desenvolvendo Inteligência Artificial?

“Only **22%** of AI professionals globally are female,

compared to **78%** who are male.”

(The Global Gender Gap Report 2018 - p.28)

# DIVERSIDADE



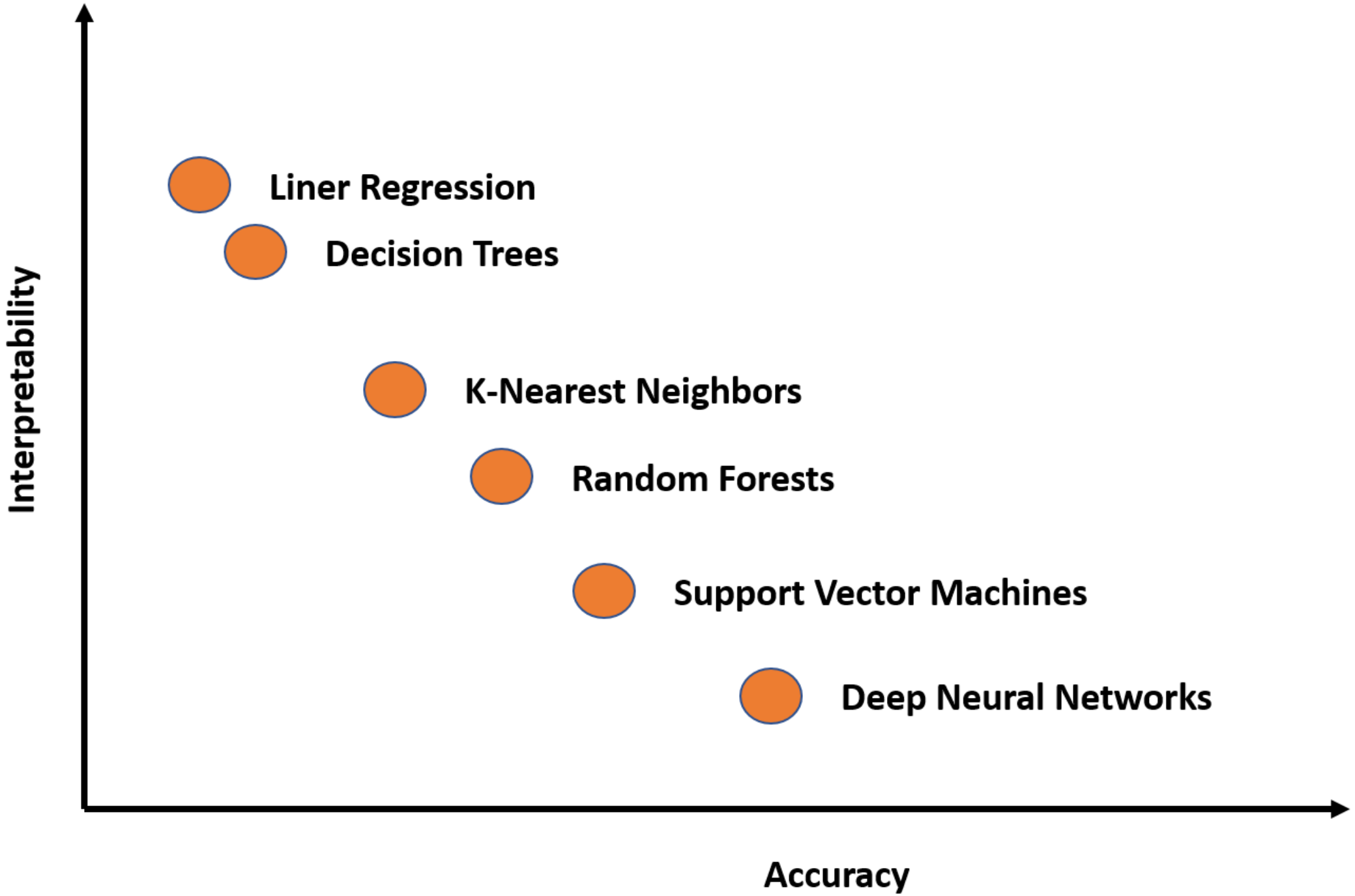
A inteligência artificial precisa aprender com o mundo real.

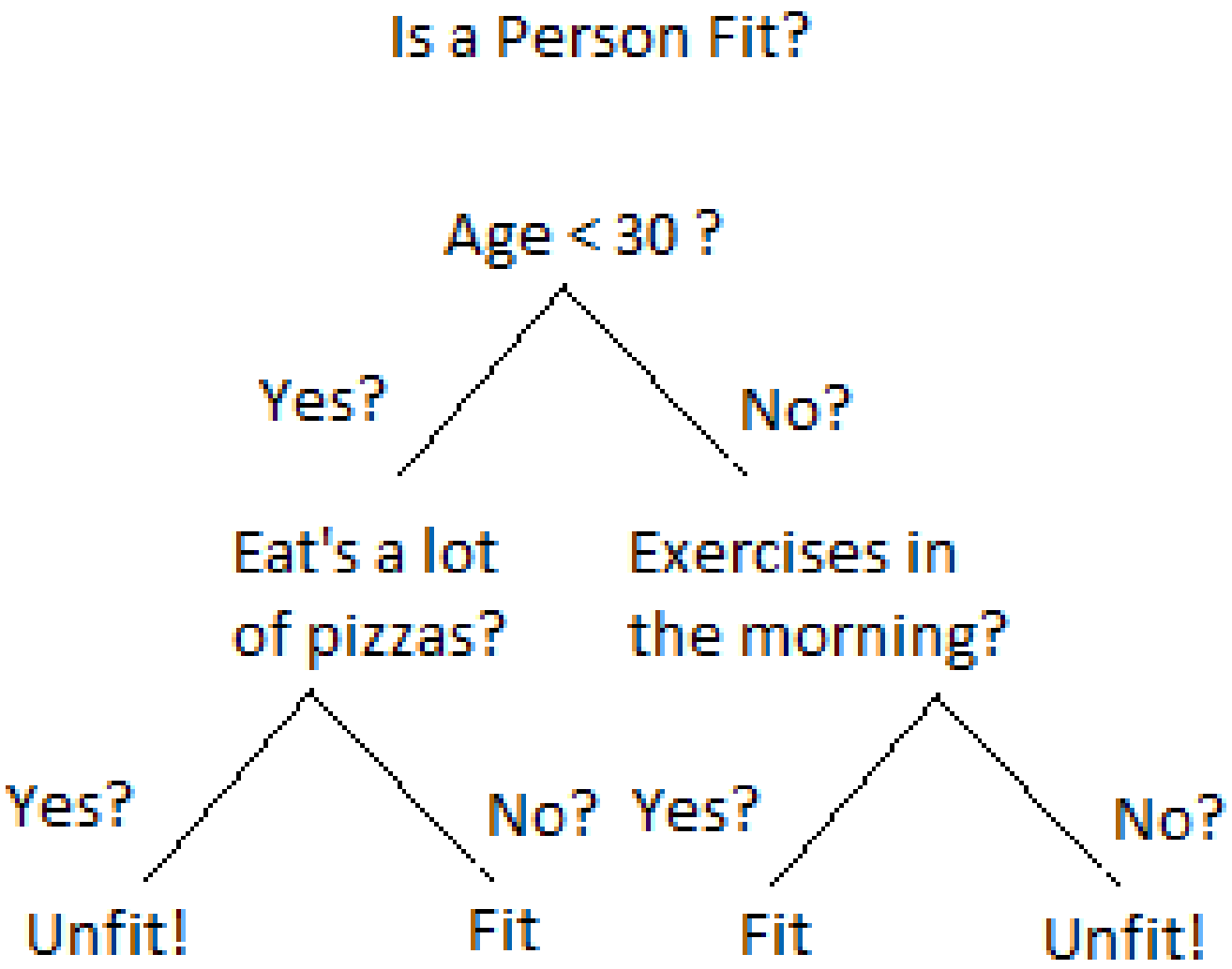
Não basta criar um computador inteligente, é preciso ensinar a ele as coisas certas.

https://about.google/stories/gender-balance-diversity-important-to-machine-learning/?hl=pt-BR

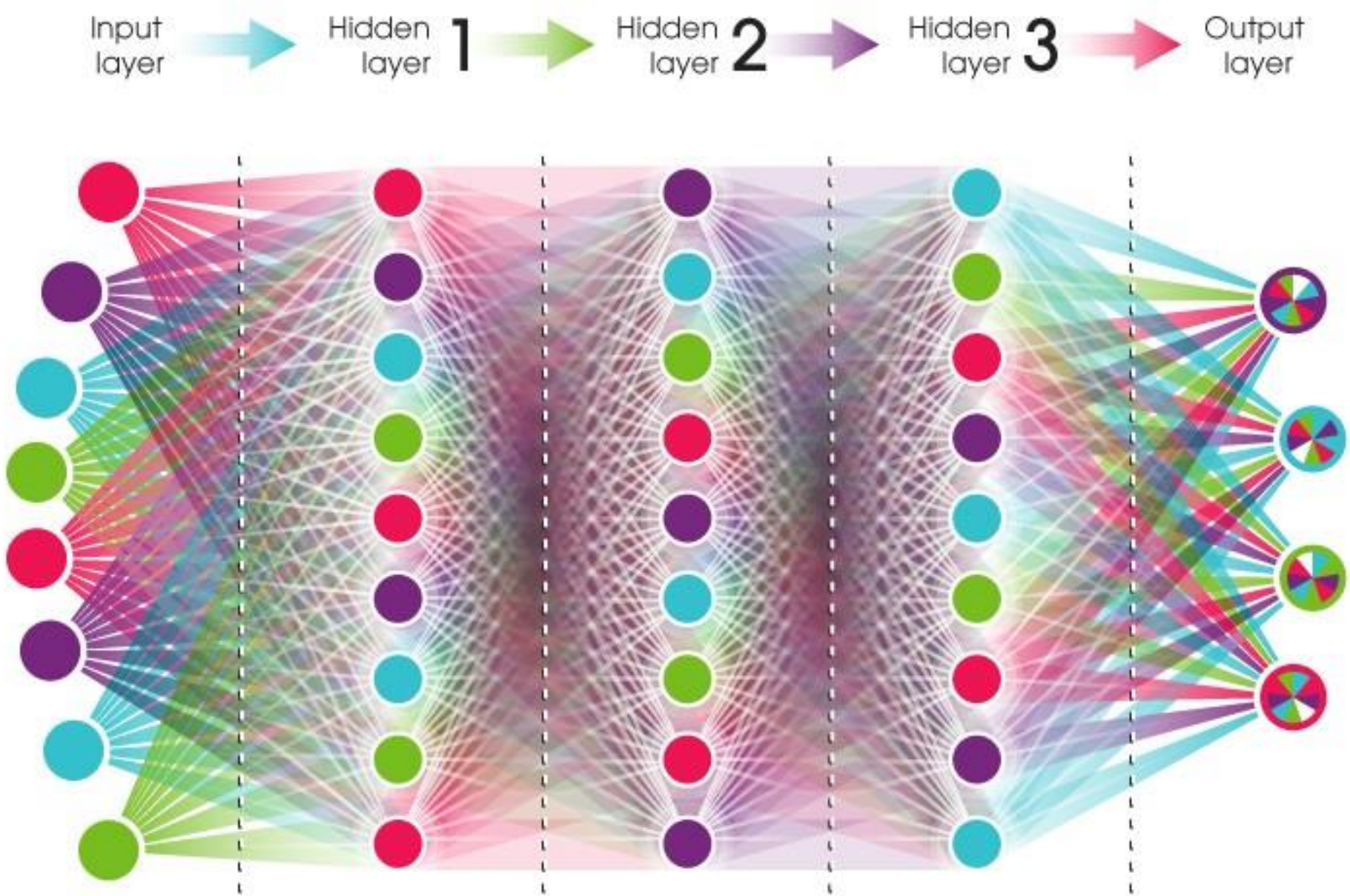

https://www.research.ibm.com/artificial-intelligence/trusted-ai/diversity-in-faces/

Estes casos ilustram um problema maior: os algoritmos de I.A. são uma **caixa-preta**, opaca e cheia de segredos.


Interpretability
Liner Regression
Decision Trees
K-Nearest Neighbors
Random Forests
Support Vector Machines
Deep Neural Networks
Accuracy


Is a Person Fit?
Age < 30 ?
Yes?
No?
Eat's a lot of pizzas?
Exercises in the morning?
Yes?
No?
Yes?
No?
Unfit!
Fit
Fit
Unfit!

# DEEP NEURAL NETWORK




INPUT
CAIXA PRETA
OUTPUT


INPUT
CAIXA
RETA
OUTPUT

# Por que abrir a caixa preta?

viés

ética

privacidade

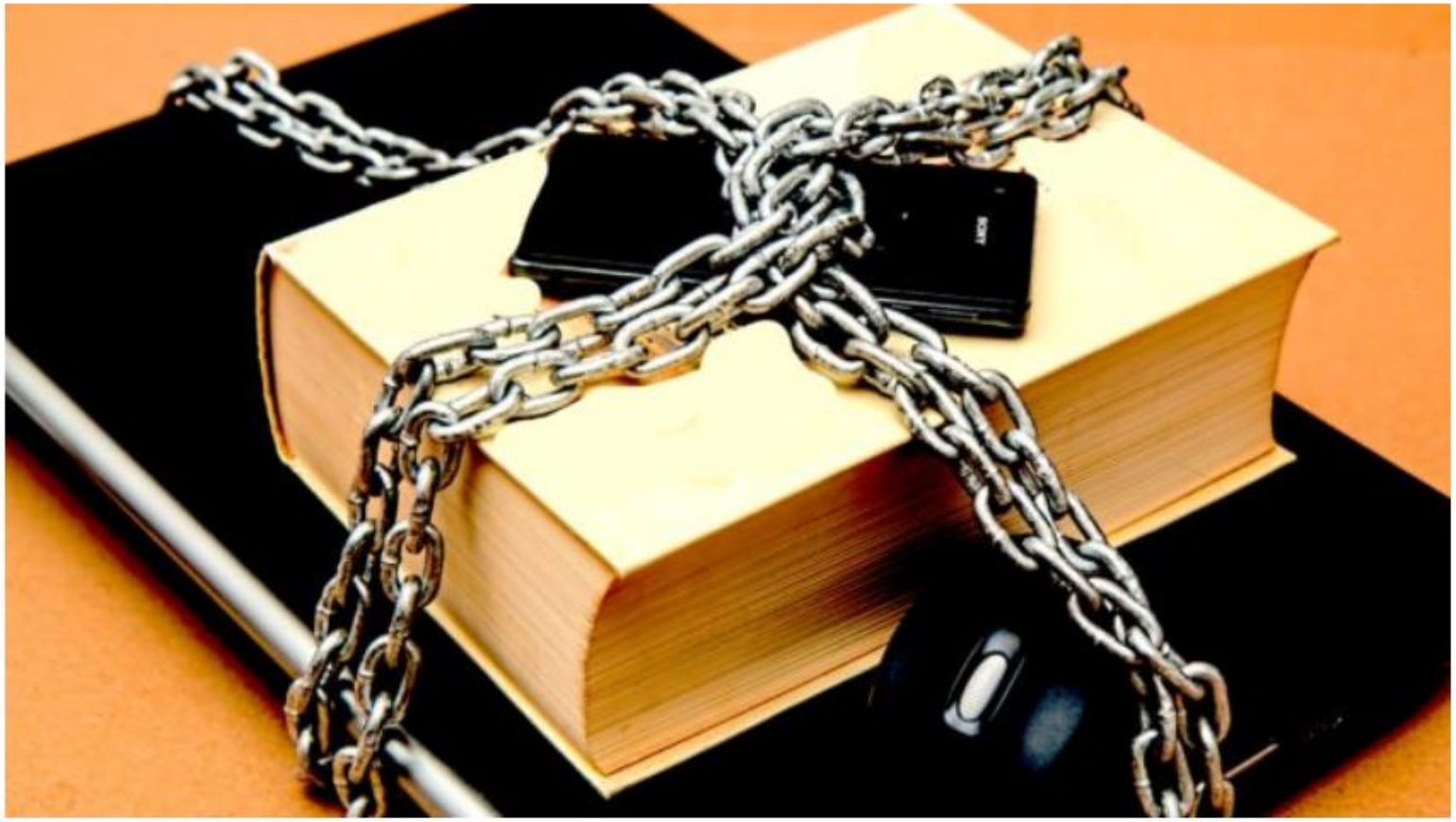
dados

legislação

# França (2019)

## France Bans Judge Analytics, 5 Years In Prison For Rule Breakers

4th June 2019 artificiallawyer Litigation Prediction 36



"Esse tipo de lei é uma desgraça para uma democracia. A Justiça é usada em nome do povo, tentar esconder informações de agentes da lei ou de cidadãos nunca será a coisa certa a fazer."

**Louis Larret Chahine**

Co-fondateur de PREDICTICE

# Estados Unidos (2019)

CNN BUSINESS

## San Francisco just banned facial-recognition technology

By Rachel Metz, CNN Business

Updated 2315 GMT (0715 HKT) May 14, 2019

Brasil
(2019)

# Hering terá que explicar o que faz com dados de reconhecimento facial de clientes

Governo instaurou processo contra a marca por indícios de práticas abusivas

**O Globo**
02/09/2019 - 13:01 / Atualizado em 02/09/2019 - 17:04

“ (...) diferentemente do que foi apontado, não realiza reconhecimento facial, mas, sim, detecção facial, por meio do qual estima apenas o gênero, a faixa etária e o humor dos consumidores, de forma anônima".

**Gerente de Marca**

Hering

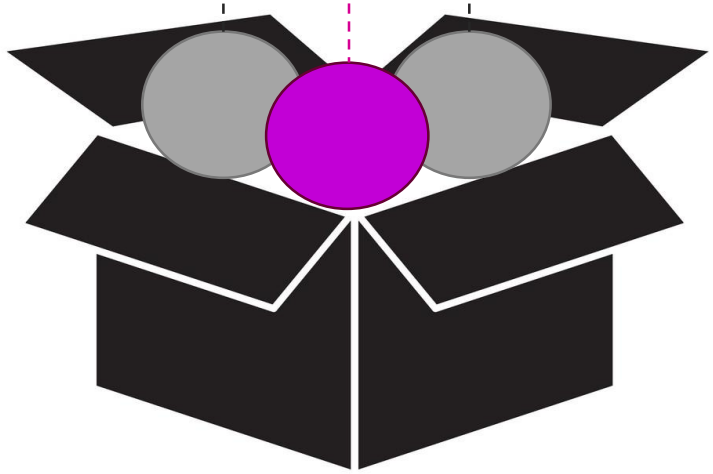
# Como abrir a caixa preta?

**EXPLICABILIDADE**

Entender a lógica por trás de cada decisão

**TRANSPARÊNCIA**

**CONFIANÇA**

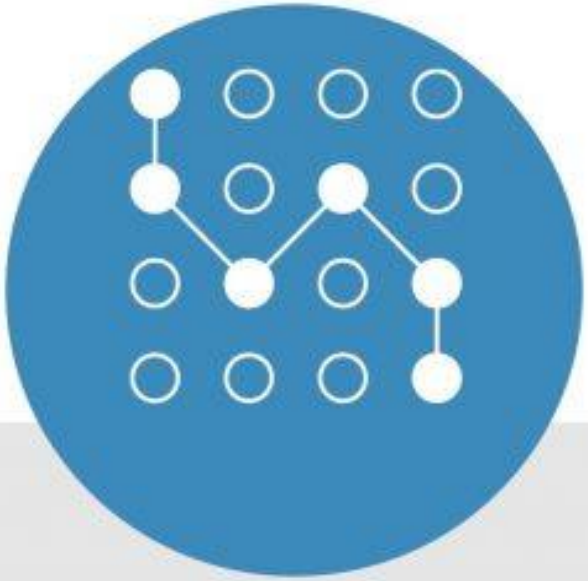


# explicabilidade

Refere-se à capacidade do sistema de explicar porque chegou a determinado resultado em linguagem compreensível para um ser humano.

## Dados explicáveis

Quais os dados utilizados para treinar o modelo e por quê?

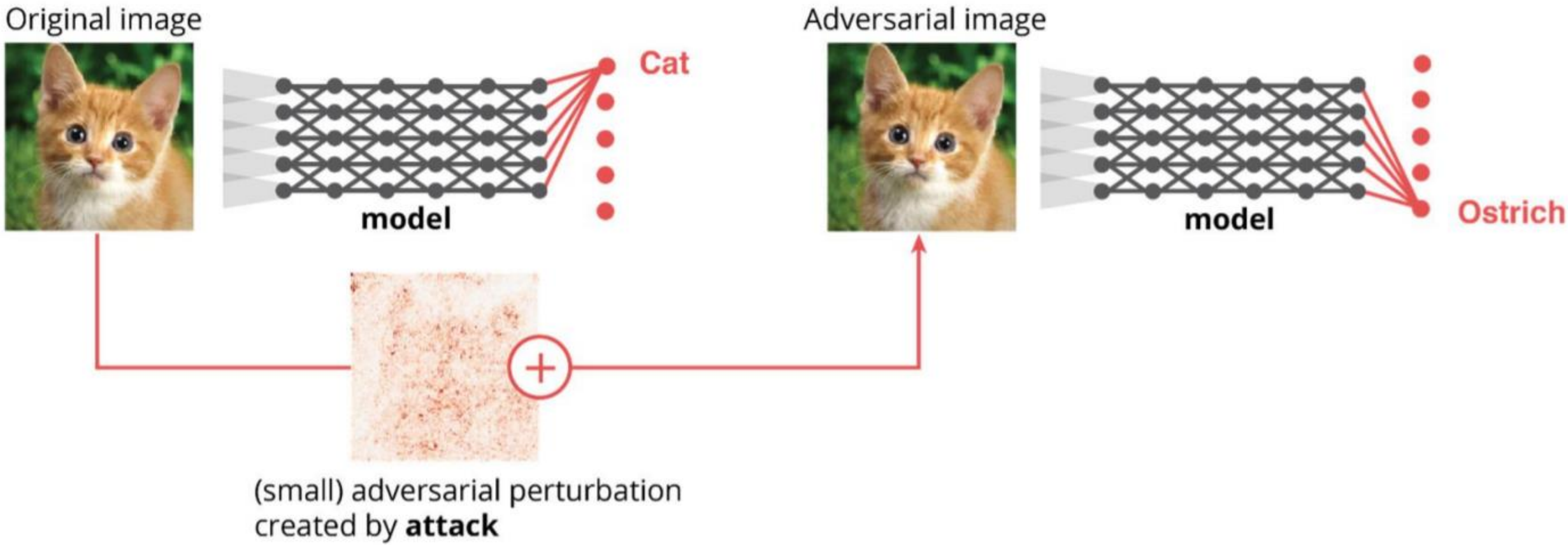


## Predições explicáveis

Quais as *características* e *pesos* utilizados para essa predição?

## Algoritmos explicáveis

Quais são as camadas e processos internos desse algoritmo?


Original image
Cat
model
Adversarial image
model
Ostrich
(small) adversarial perturbation
created by **attack**

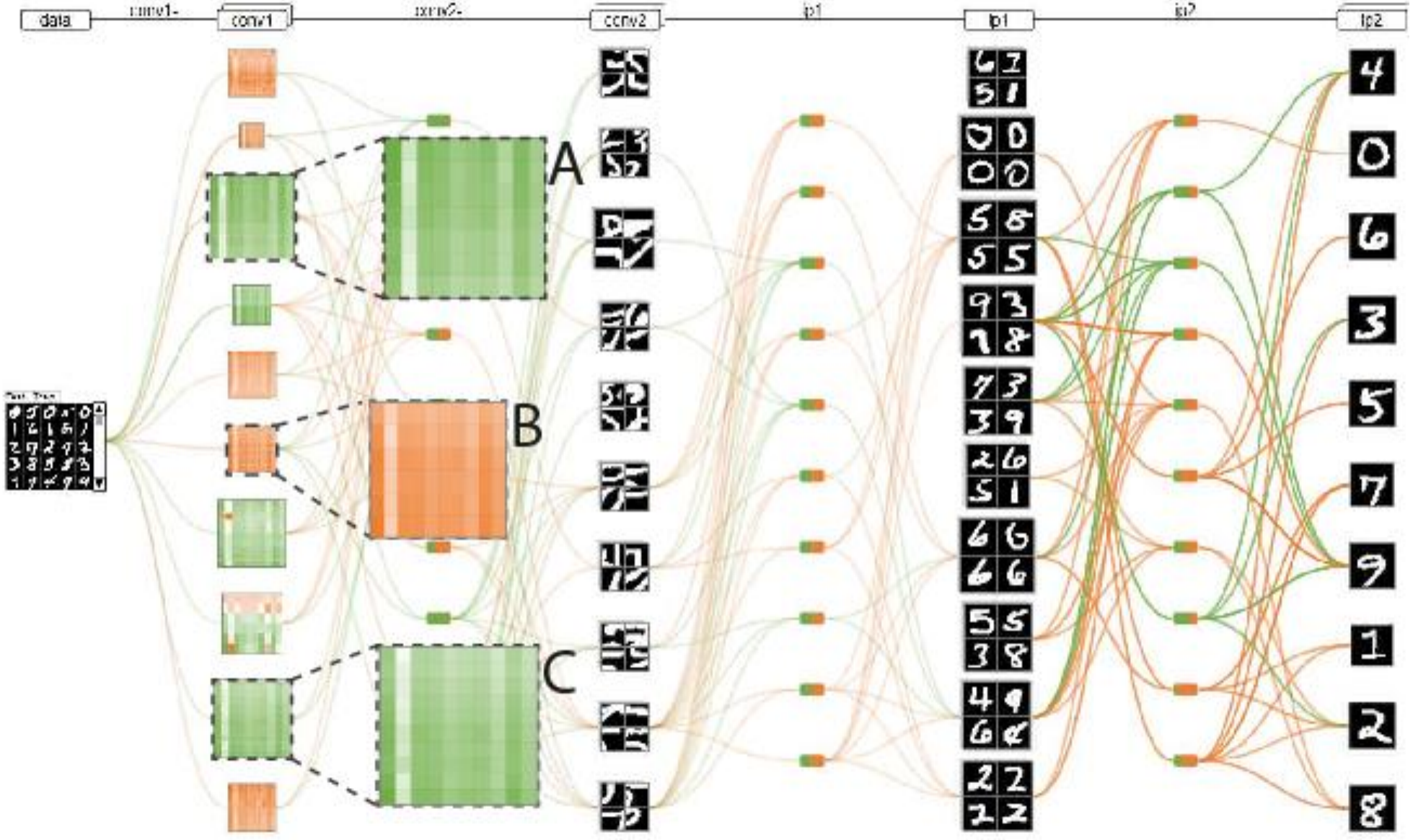



*Figure 5.* Neurons in a model that is too wide. Many neurons in

As soluções de Inteligência Artificial não são e não serão infalíveis.

Mas, a explicabilidade pode ajudar...

“O sucesso na criação da IA será o maior acontecimento na história da humanidade. Infelizmente, também poderá ser o último, a menos que aprendamos como evitar os riscos.”

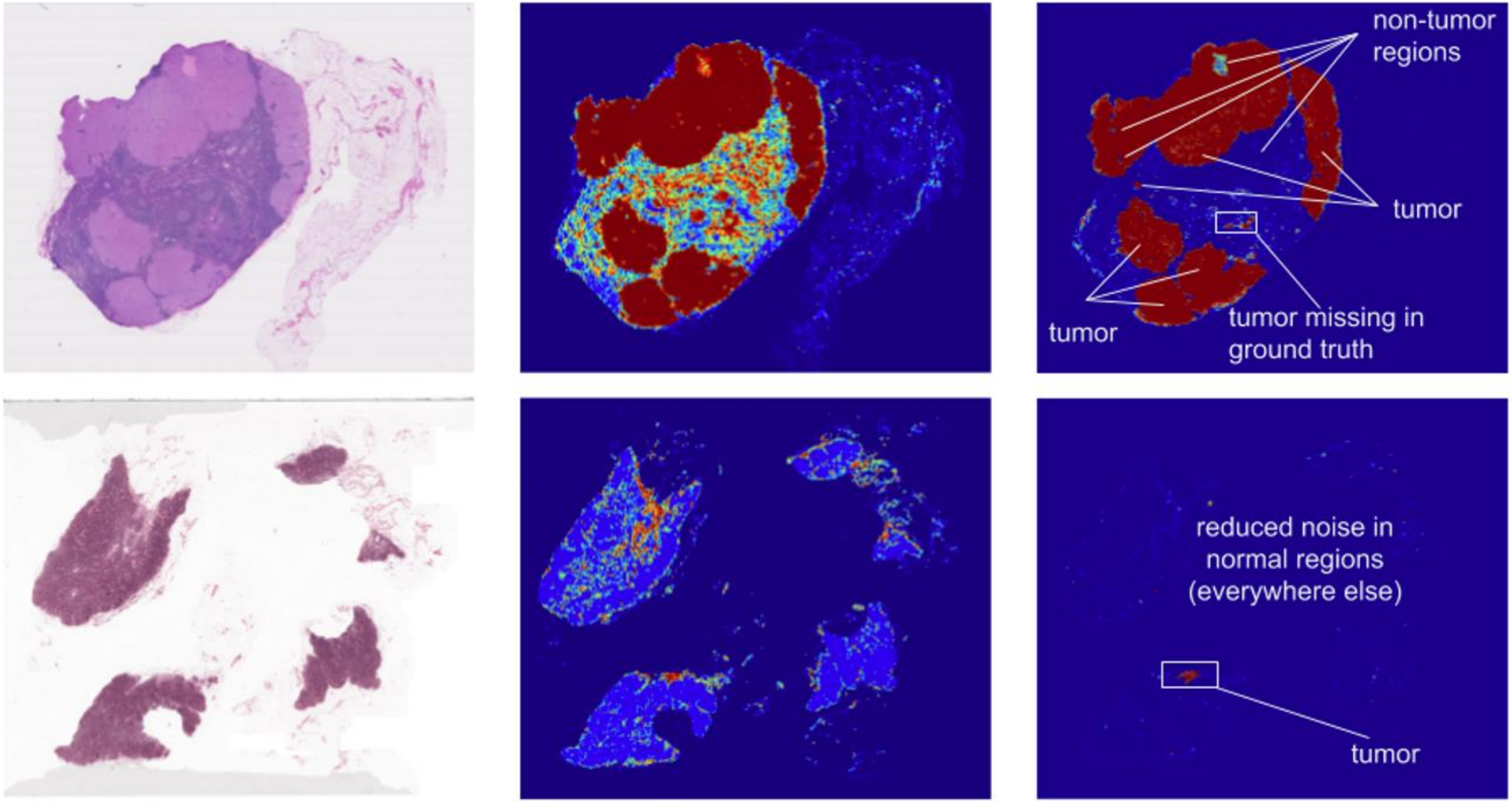
**Stephen Hawking**

McKinsey Global Institute

# Applying artificial intelligence for social good

November 2018 | Discussion Paper

# OBJETIVOS DE DESENVOLVIMENTO SUSTENTÁVEL


non-tumor regions
tumor
tumor
tumor missing in ground truth
reduced noise in normal regions (everywhere else)
tumor
1
Tumor probability
0

# Portal da Transparência

CONTROLADORIA-GERAL DA UNIÃO



## Dados abertos

Aqui é possível baixar os dados apresentados no Portal da Transparência do Governo Federal, em formato aberto, possibilitando que os usuários façam cruzamentos e análises específicas, de acordo com suas necessidades.

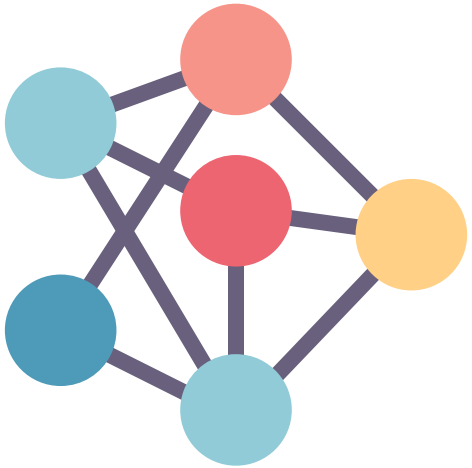
Os arquivos são disponibilizados em formato CSV (clique aqui para mais informações).

- ORÇAMENTO PÚBLICO
- DESPESAS PÚBLICAS
- CARTÃO DE PAGAMENTO
- RECEITAS PÚBLICAS
- LICITAÇÕES E CONTRATOS
- CONVÊNIOS E INSTRUMENTOS CONGÊNERES
- BENEFÍCIOS AO CIDADÃO

http://www.portaltransparencia.gov.br/download-de-dados

# Desafios

Acessibilidade de dados



Explicabilidade dos algoritmos

Diversidade nos talentos


https://serenata.ai/


https://brasil.io/home


https://colaboradados.github.io/

Se a tecnologia quiser ajudar na construção de uma sociedade mais justa, ela tem que ser aberta e transparente.

Antes de falar sobre futuro...

... precisamos falar sobre o que está acontecendo hoje, agora.

# Obrigada!

Carla Vieira

@carlaprvieira

carlaprv@hotmail.com

bit.ly/goias-carla

# Referências


- Relatórios do AI NOW
- Racial and Gender viés in Amazon Rekognition
- Diversity in faces (IBM)
- Google video – Machine Learning and Human viés
- Visão Computacional e Vieses Racializados
- Estudo Machine viés on Compas
- Machine Learning Explainability Kaggle
- Predictive modeling: striking a balance between accuracy and interpretability

# Referências


- Racismo Algorítmico em Plataformas Digitais: microagressões e discriminação em código
- Metrics for Explainable AI: Challenges and Prospects
- The Mythos of Model Interpretability
- Towards Robust Interpretability with Self-Explaining Neural Networks
- The How of Explainable AI: Post-modelling Explainability