GAN, SEGURANÇA, VIÉS E ÉTICA

Lucas Bastos
Gabriel Tormin
Marco Aurélio
Filipe Ribeiro

20 de maio de 2021

UnB - Departamento de Estatística

GAN

G - Generative
A - Adversarial
N - Network

Redes Adversárias Generativas

GAN

Duas redes competindo:

Generativa *vs* Discriminativa

Introduzido por Ian Goodfellow et al. (2014)

GAN

Podemos pensar em algoritmos generativos como o oposto dos algoritmos discriminativos.

GAN

Podemos pensar em algoritmos generativos como o oposto dos algoritmos discriminativos.

Discriminativos reduzem a informação para obter rótulos.

GAN

Podemos pensar em algoritmos generativos como o oposto dos algoritmos discriminativos.

Discriminativos reduzem a informação para obter rótulos.

Generativos criam informação para um dado rótulo.

GAN

Exemplo: classificação de e-mails em *spam* ou não *spam*.

GAN

Exemplo: classificação de e-mails em *spam* ou não *spam*.

Discriminativas: $p(x|y)$

GAN

Exemplo: classificação de e-mails em *spam* ou não *spam*.

Discriminativas: $p(x|y)$
Generativo: Como obter x? Otimizando $p(x|y)$

GAN

Loop de feedback duplo:

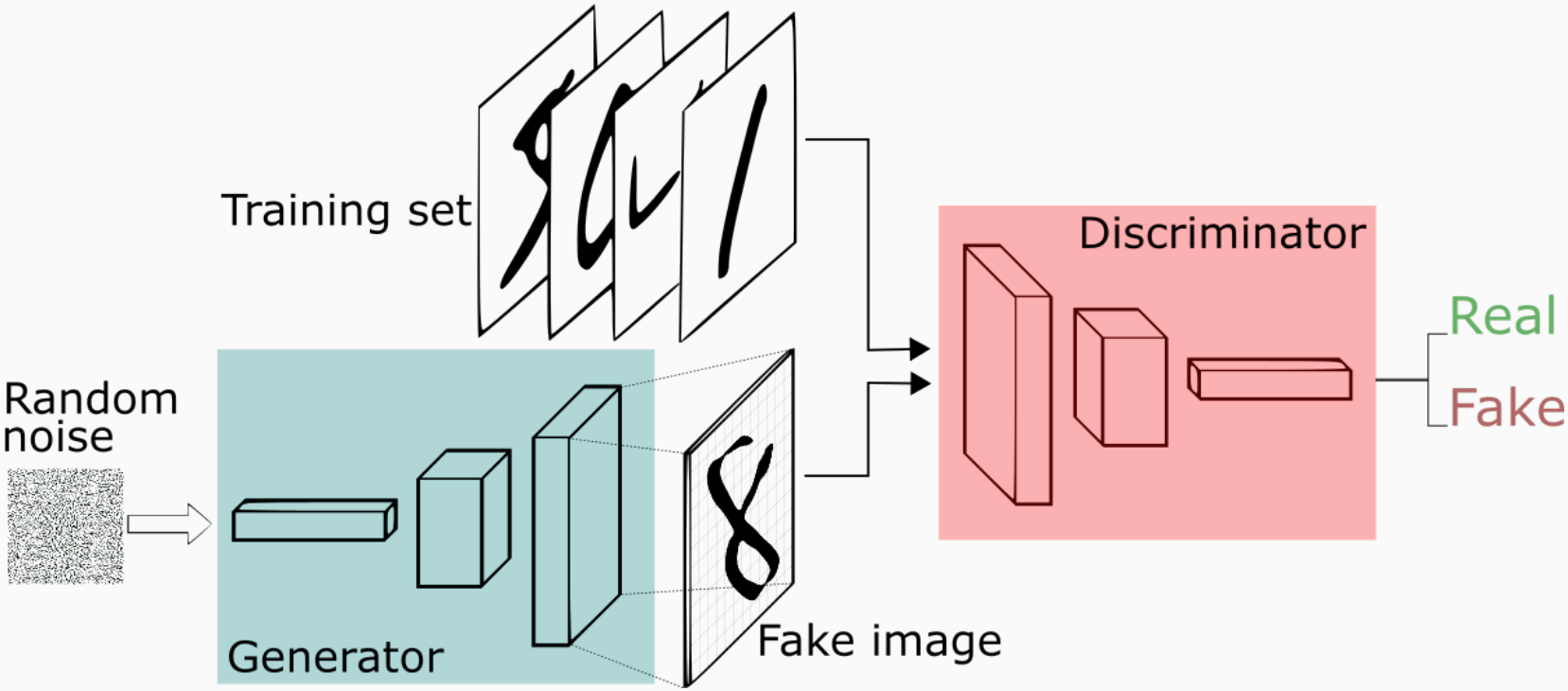


Figura: deeplearningbook.com.br - Capítulo 54

GAN

GANs podem aprender qualquer distribuição.

Imagens, música, fala, prosa, etc.

Combinando esses, pode gerar Deep Fakes.

Contudo, em 2019 DeepMind mostrou que os Autoencoders Variacionais (VAEs) poderiam superar as GANs na geração de faces.

GAN

Figura: Two of Barrat's GAN-generated nudes. Images by Robbie Barrat

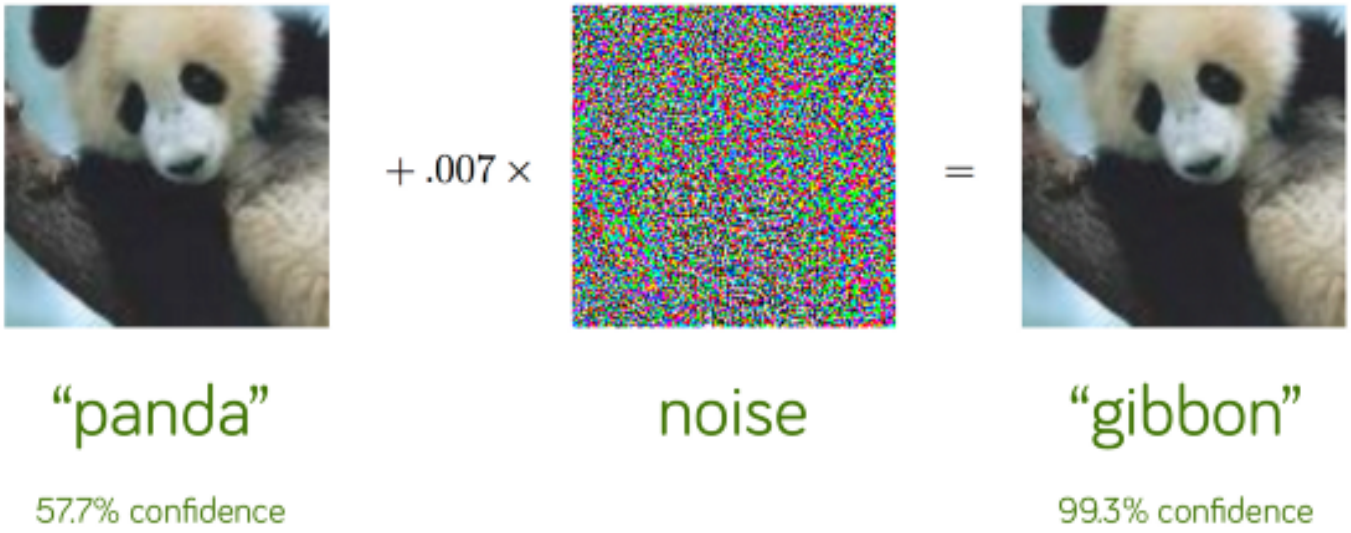
GAN

Mais exemplos:

Figura: Fake Obama - BBC

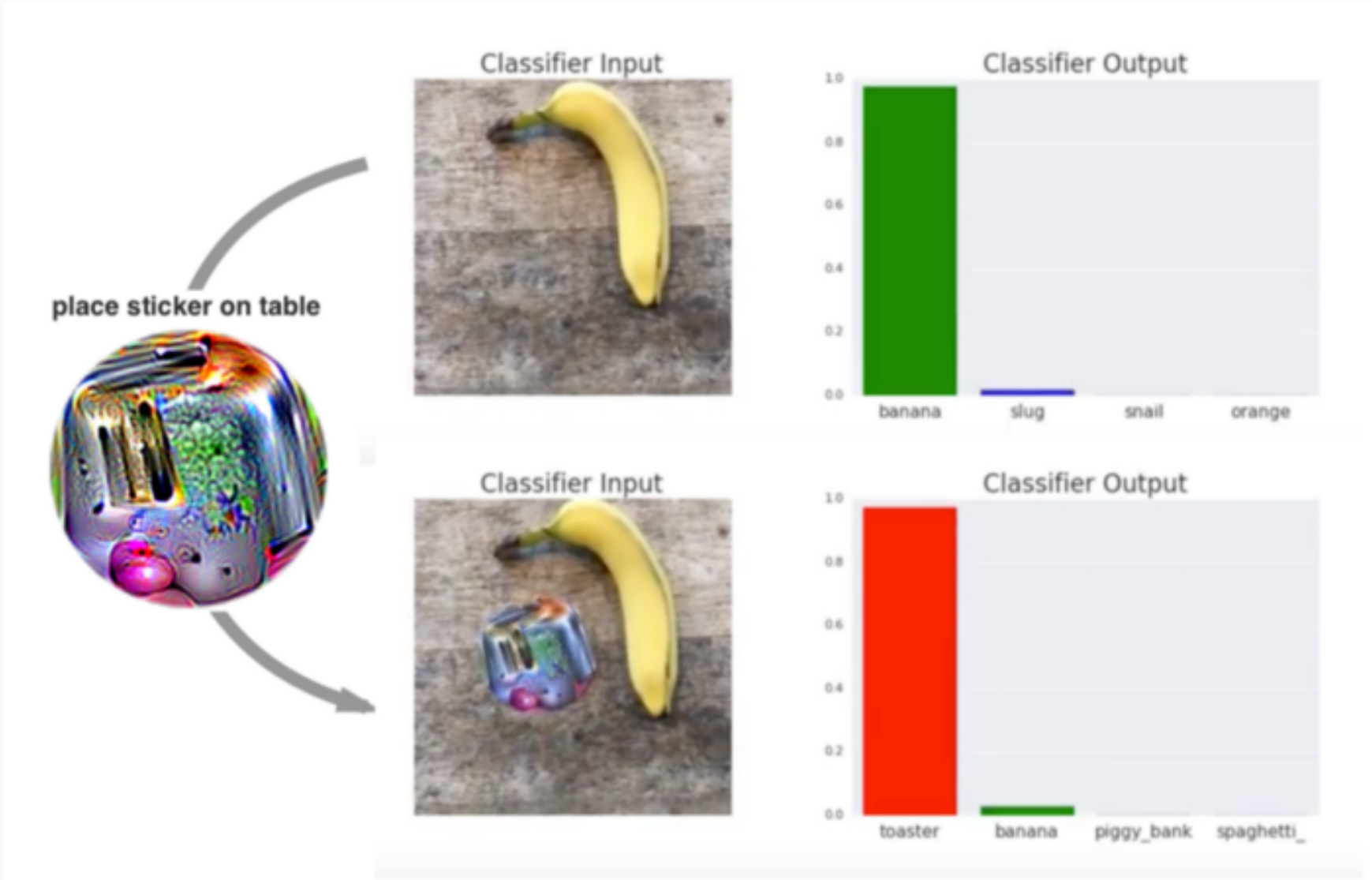
Figura: GANs interativas Nvidia

Ataques adversariais



Fonte: https://towardsdatascience.com/adversarial-attacks-in-machine-learning-and-how-to-defend-against-them.

Ataques adversariais



Fonte: https://towardsdatascience.com/adversarial-attacks-in-machine-learning-and-how-to-defend-against-them.

Ataques adversariais

Como lidar com esses ataques?

- Denoising;
- Inserção de ataques para que a rede seja treinada a identificá-los;
- BaRT (Barrage of Random Transforms).

Perigos:

- roubo de identidade em sistemas de reconhecimento facial;
- sinais de trânsito usados por veículos autônomos;
- detectores de notícias falsas;

...

Segurança, Viés e Ética

- Dados históricos e seu viés intrínseco

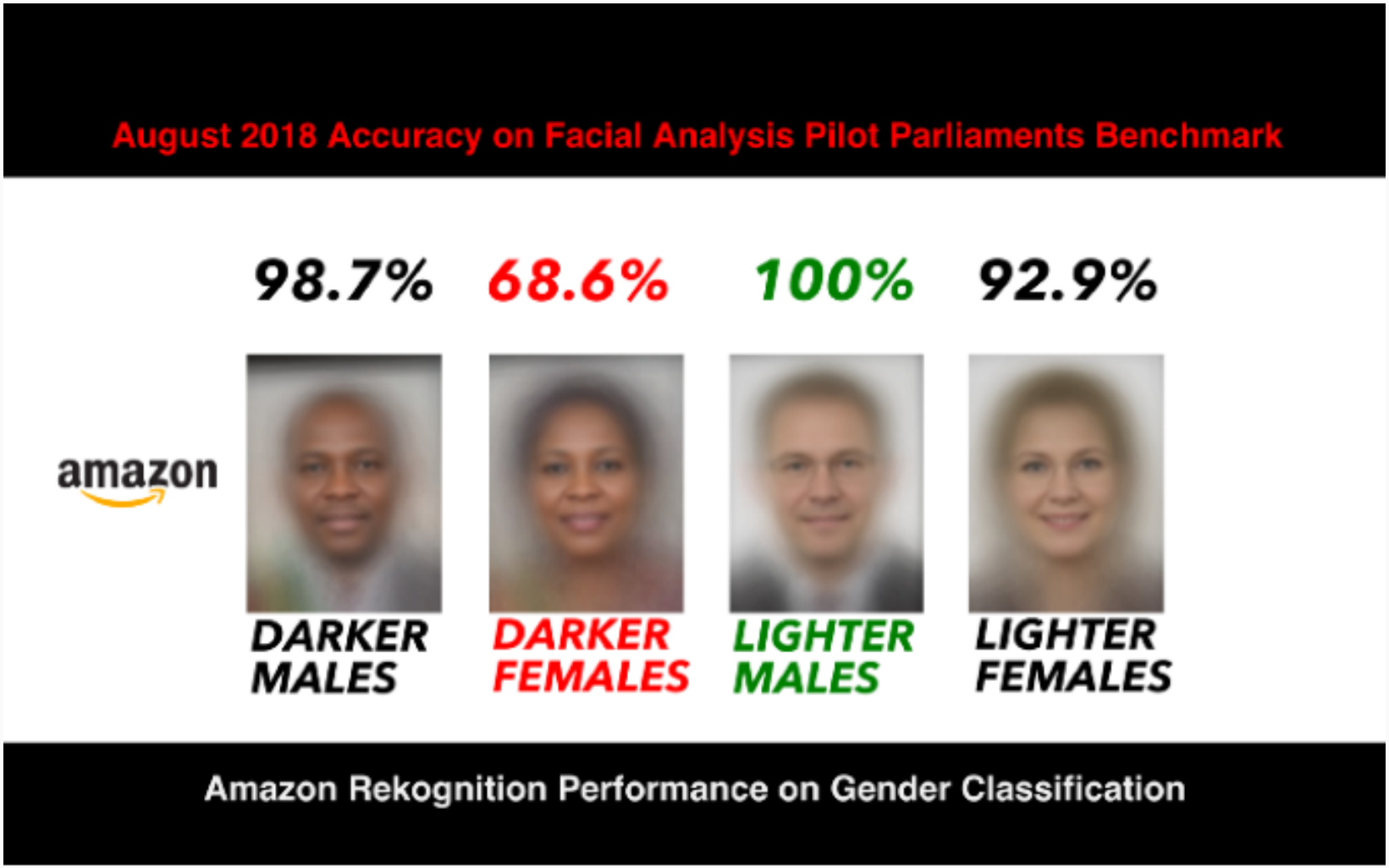


Figura: Fonte: Joy Buolamwini - Racial and Gender bias in Amazon Rekognition — Commercial AI System for Analyzing Faces.

Segurança, Viés e Ética

- Viés na avaliação de crédito pelos bancos: taxa de crédito negado para negros nos EUA de 25%. Para brancos, 10%.

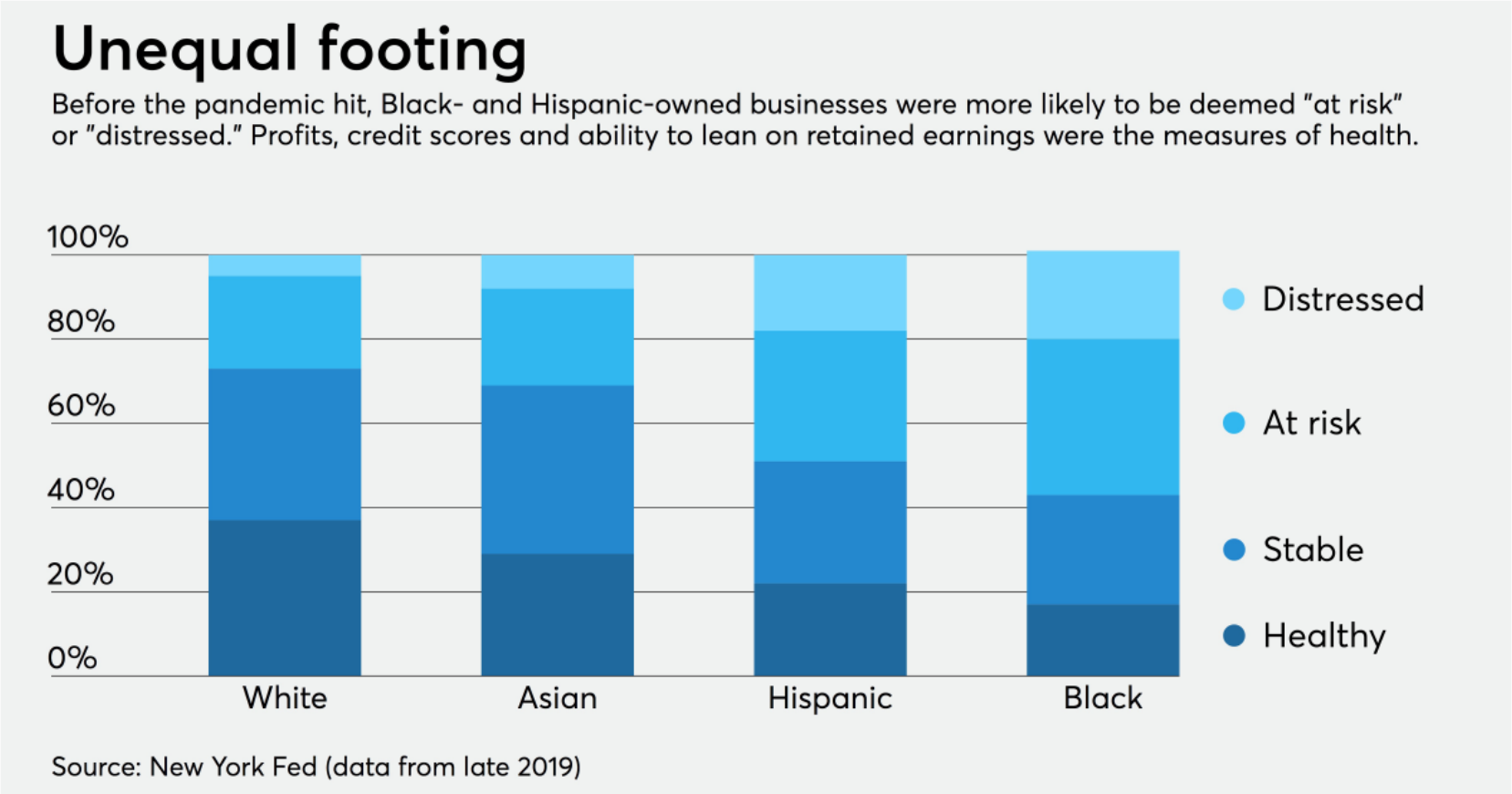


Figura: Fonte: A House Divided - How Race Colors the Path to Homeownership.

Segurança, Viés e Ética

Avaliação de professores do sistema público de ensino estadunidense a fim de obterem estabilidade.

- Caso: Daniel Santos, professor premiado e muito bem avaliado em diversas escolas em que lecionou, não era considerado apto pelo algoritmo.

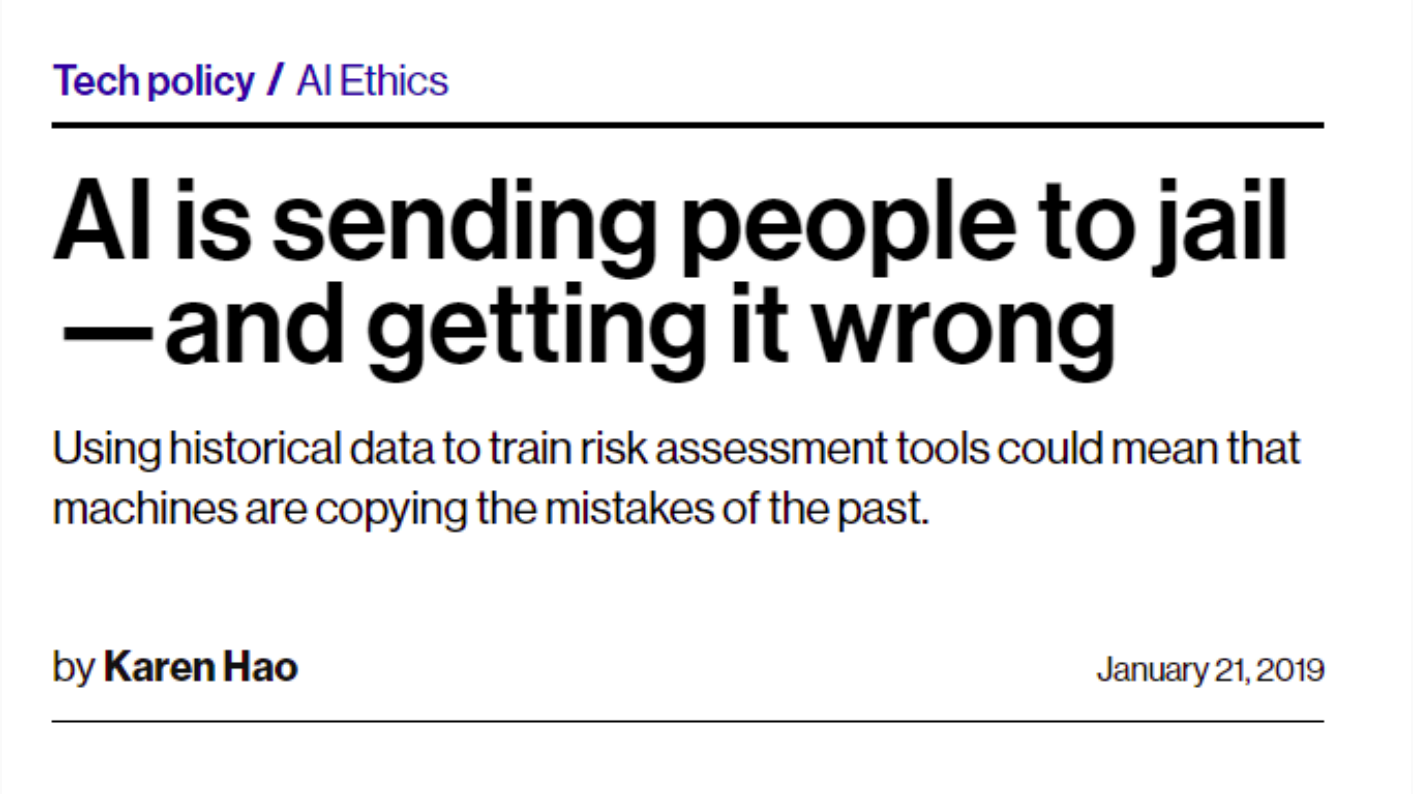
Fonte:https://www.chron.com/news/houston-texas/education/article/Houston-ISD-settles-with-union-over-teacher-12267893.php

Segurança, Viés e Ética

Avaliação de detentos em liberdade provisória nos EUA.

- O algoritmo utilizado pelo Departamento de Justiça tem a última palavra acima do juiz e do oficial de condicional.

- ”Profecia autorrealizada”



Tech policy / AI Ethics

AI is sending people to jail —and getting it wrong

Using historical data to train risk assessment tools could mean that machines are copying the mistakes of the past.

by **Karen Hao** January 21, 2019

Fonte: MIT Technology Review

Segurança, Viés e Ética

- Uso de dados pessoais e a violação da privacidade.

Segurança, Viés e Ética

No Reino Unido, o reconhecimento facial de pedestres identificou milhares de pessoas erroneamente. Em Wales, o índice de falha chegou a 92%.

Fonte: https://www.lbc.co.uk/news/police-use-of-facial-recognition-technology-ad-hoc/

Segurança, Viés e Ética

- Devido Processo Legal;
- Contraditório e Ampla Defesa;
- Inviolabilidade da pessoa humana;
- Princípio da não-autoincriminação;

...

Segurança, Viés e Ética

- Desinformação e manipulação de massa.

https://www.cleveland.com/darcy/2017/07/trump-putin$_{meeting}$$_{accounts}$.html

Segurança, Viés e Ética

- Remodelagem comportamental

Fonte:https://thehackernews.com/2016/03/tay-artificial-intelligence.html

Segurança, Viés e Ética

- Escore de crédito social.

https://www.verdict.co.uk/facebook-rating-score-china-social-credit/

Segurança, Viés e Ética

- Automação de armas e eficiência bélica.

https://digitalbusiness.law/2017/10/the-use-of-ai-in-weapons-systems-the-uk-and-us-legal-and-regulatory-framework/