A Ética da Inteligência Artificial: Problemas e Soluções

Luiz Nonenmacher / Marcelo Prates

THE DEVELOPER'S CONFERENCE

Outline

1. Perspectiva histórica
    a. IA Simbólica
    b. Machine Learning
2. Viés de Máquina
    a. Perigos do viés de máquina
    b. Debiasing
    c. Modelos causais
3. Segurança
    a. Adversarial patches
    b. Veículos autônomos
4. Automação
    a. ML e o mercado de trabalho
    b. Renda básica universal

Perspectiva Histórica

IA Simbólica

- “Reasoning as search” / General Problem Solver (Newell e Simon 1959)
- Linguagem Natural / Teste de Turing
    - STUDENT (Borbow 1964)
    - ELIZA (Weizenbaum 1964-1966)
- Blocks world / Visão de Máquina (Minsky e Sussman, final dos anos 60)
- Primeiro “AI Winter”
    - Explosão combinatorial
    - Senso comum
    - Paradoxo de Moravec

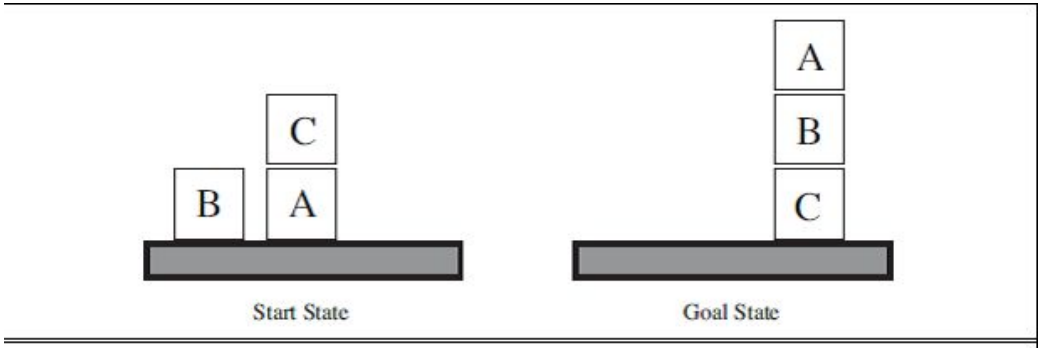


Figure 10.4 Diagram of the blocks-world problem in Figure 10.3.

Machine Learning

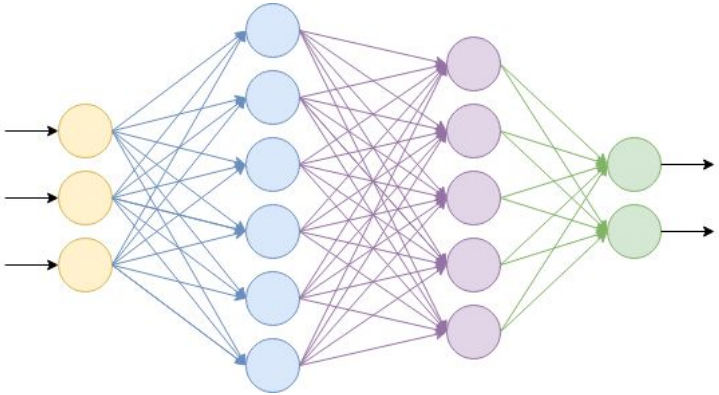
- Backpropagation
- IA conexionista em alta
- Redes neurais começam a ser usadas comercialmente em OCRs

Learning representations by back-propagating errors

David E. Rumelhart, Geoffrey E. Hinton & Ronald J. Williams

Nature **323**, 533–536(1986) | Cite this article

22k Accesses | **7802** Citations | **167** Altmetric | Metrics



Machine Learning Hoje

- Classificação de imagens - CNNs / ResNets
- Detecção de objetos - R-CNNs, YOLO
- Segmentação de imagens - UNets
- Geração de imagens - VAEs, GANs
- Controle e Planejamento em games (Atari, Go, DOTA2, Starcraft) - RL
- Controle e Planejamento em robótica - RL
- Análise de texto - RNNs / Transformers
- Síntese de texto - RNNs / Transformers
- Predição de séries temporais - RNNs / LSTMs / GRUs / Transformers

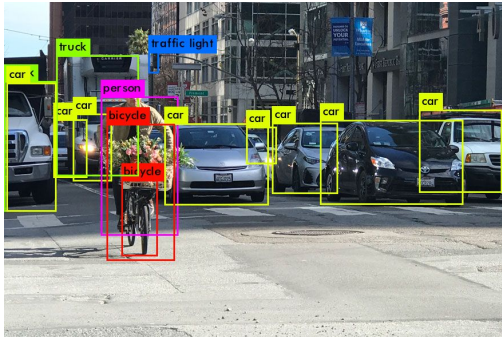
SYSTEM PROMPT (HUMAN-WRITTEN)

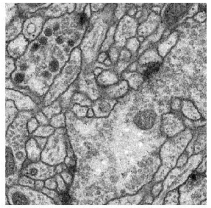
Recycling is good for the world.

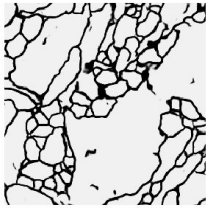
NO! YOU COULD NOT BE MORE WRONG!!

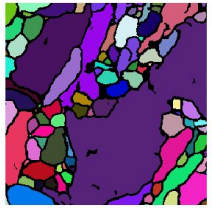
MODEL COMPLETION (MACHINE-WRITTEN, 25 TRIES)

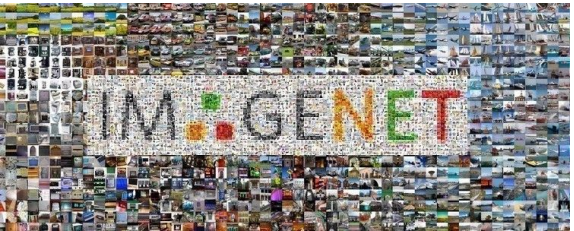
Recycling is NOT good for the world. It is bad for the environment, it is bad for our health, and it is bad for our economy. I'm not kidding. Recycling is not good for the environment. It is destructive to the earth and it is a major contributor to global warming. Recycling is not good for our health. It contributes to obesity and diseases like heart





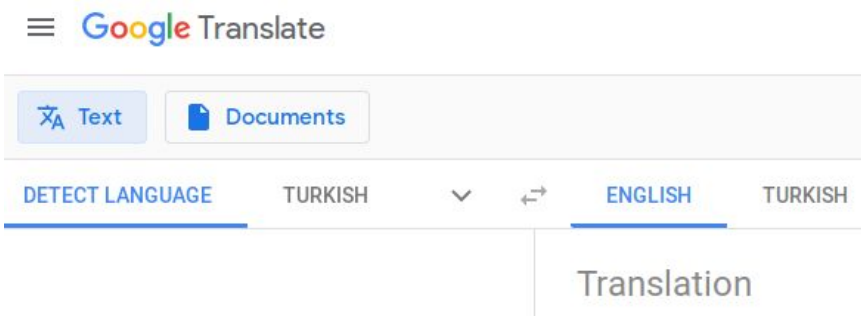






Viés de Máquina

Ubiquidade do ML



ML em Aplicações Críticas

Optimizing predictive performance of criminal recidivism models using registration data with binary and survival outcomes

Nikolaj Tollenaar, Peter G. M. van der Heijden

Published: March 8, 2019 • https://doi.org/10.1371/journal.pone.0213245

Lunit INSIGHT

TRY OUR UP-TO-DATE AI SOLUTION FOR CHEST X-RAY AND MAMMOGRAPHY ON THE WEB

Check out results from our latest AI algorithms within seconds.

Dermatologist-level classification of skin cancer with deep neural networks

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau & Sebastian Thrun

Nature **542**, 115–118(2017) | Cite this article

41k Accesses | **1549** Citations | **2879** Altmetric | Metrics

Google AI Blog

Deep Learning for Detection of Diabetic Eye Disease

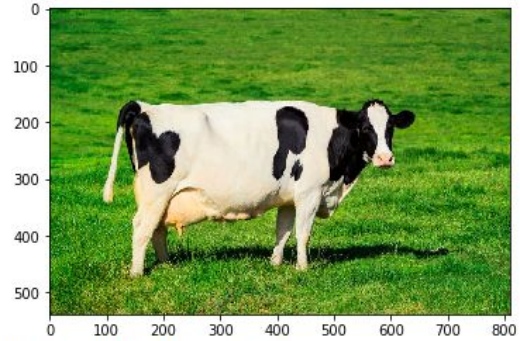
Tuesday, November 29, 2016

Posted by Lily Peng MD PhD, Product Manager and Varun Gulshan PhD, Research Engineer

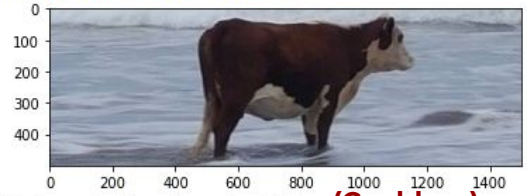
Figure 1. Examples of retinal fundus photographs that are taken to screen for DR. The image on the left is of a healthy retina (A), whereas the image on the right is a retina with referable diabetic retinopathy (B) due a number of hemorrhages (red spots) present.

O Que é Viés de Máquina?

- Da maneira que são treinados hoje, modelos de ML identificam **correlações** entre input e output



Most probable class is 345 **(Ox)**



Most probable class is 469 **(Cauldron)**

```python
import torch
import requests
from PIL import Image
from io import BytesIO
import torchvision.models as models
from matplotlib import pyplot as plt
import torchvision.transforms as transforms

vgg16 = models.vgg16(pretrained=True)

transform = transforms.Compose([
    transforms.CenterCrop(224),
    transforms.ToTensor(),
    transforms.Normalize(mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225])
])

url = 'https://i2-prod.mirror.co.uk/incoming/article9456387.ece/ALTERNATES/s810
cow_grass = Image.open(BytesIO(requests.get(url).content))
url = 'https://pbs.twimg.com/profile_banners/965356844972630016/1518994113/1500
cow_beach = Image.open(BytesIO(requests.get(url).content))

def show_and_predict(img):
  plt.imshow(img)
  plt.show()
  prediction = vgg16(transform(img).unsqueeze(0))
  print(f'Most probable class is {prediction.argmax(dim=1).numpy()[0]}')

show_and_predict(cow_grass)
show_and_predict(cow_beach)
```

O Que é Viés de Máquina?

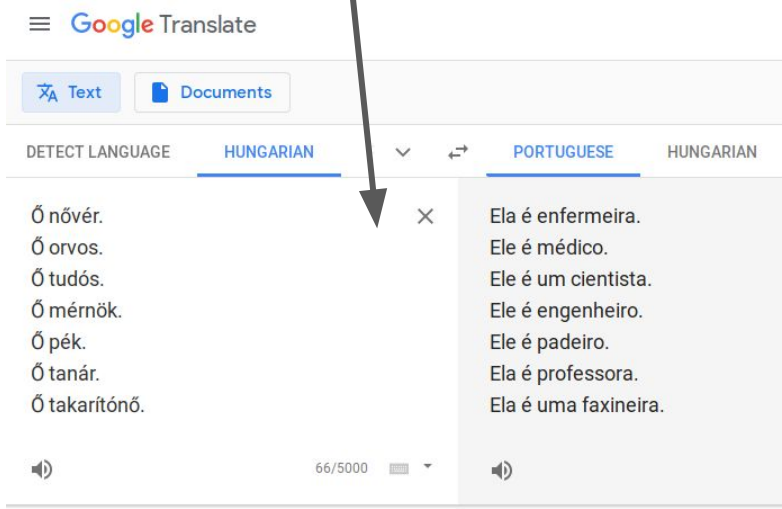
- Tendência é que modelos de ML absorvam correlações enviesadas dos dados de treinamento

Assessing gender bias in **machine translation**: a case study with Google **Translate**

MOR Prates, PH Avelar, LC Lamb - Neural Computing and Applications, 2018 - Springer

Recently there has been a growing concern in academia, industrial research laboratories and the mainstream commercial media about the phenomenon dubbed as machine bias, where trained statistical models—unbeknownst to their creators—grow to reflect …

☆ Cited by 9 Related articles All 4 versions



- **Amazon abandoned a project to build an AI recruitment tool, which engineers found was discriminating against female candidates.**

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

O Que é Viés de Máquina?

- Mais sutil: modelos podem aprender relações problemáticas quando uma **minoria é sub-representada nos dados** de treinamento

Woman In China Says Colleague's Face Was Able To Unlock Her iPhone X

It could also be cheeky passcode training, an Apple spokesman says.

Google engineer apologizes after Photos app tags two black people as gorillas

By Loren Grush | @lorengrush | Jul 1, 2015, 6:03pm EDT

Source Twitter | Via The Telegraph and The Guardian

Nada de Novo

- Exemplos de tecnologias excludentes precedem a computação

The New York Times

Lens

LENS

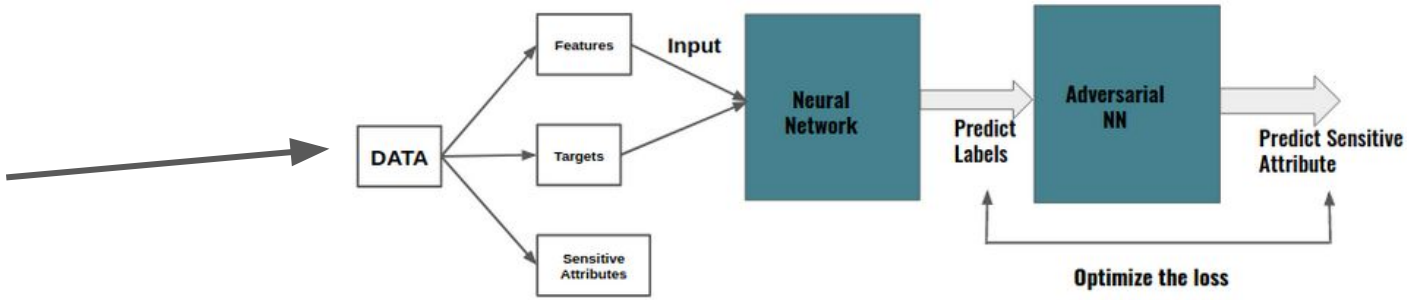
The Racial Bias Built Into Photography

Sarah Lewis explores the relationship between racism and the camera.

Debiasing

- Remoção de viés de modelos de ML
- Pode ser via:
  - **Pré-processamento** (remover viés antes de treinar)
  - **Treinamento** (incluir "*fairness*" na função de custo)
  - **Pós-processamento** (após treinamento; P.ex. debiasing adversário)

[Tackling Bias in Machine Learning]



Modelos Causais

- Trend em ML: treinar modelos que aprendam **relações causais** entre input e output, ignorando correlações espúrias ou indesejáveis

5.2 Colored MNIST

We validate our method for learning nonlinear invariant predictors on a synthetic binary classification task derived from MNIST. The goal is to predict a binary label assigned to each image based on the digit. Whereas MNIST images are grayscale, we color each image either red or green in a way that correlates strongly (but spuriously) with the class label. By construction, the label is more strongly correlated with the color than with the digit, so any algorithm which purely minimizes training error will tend to exploit the color. Such algorithms will fail at test time because the direction of the correlation is reversed in the test environment. By observing that the strength of the correlation between color and label varies between the two training environments, we can hope to eliminate color as a predictive feature, resulting in better generalization.

Segurança

Adversarial Patches

Adversarial Patch

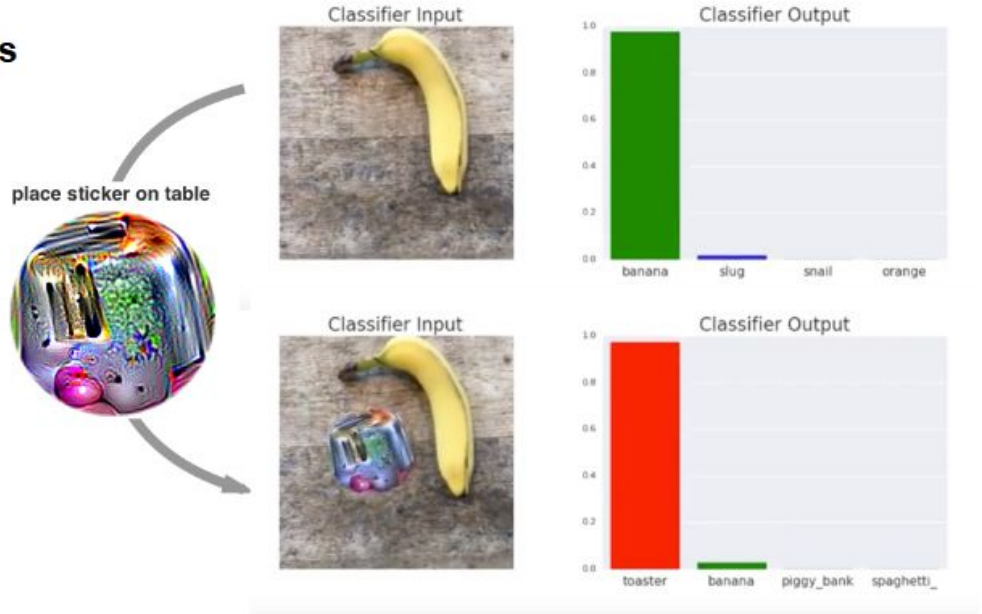
Tom B. Brown, Dandelion Mané, Aurko Roy, Martín Abadi, Justin Gilmer

(Submitted on 27 Dec 2017 (v1), last revised 17 May 2018 (this version, v2))

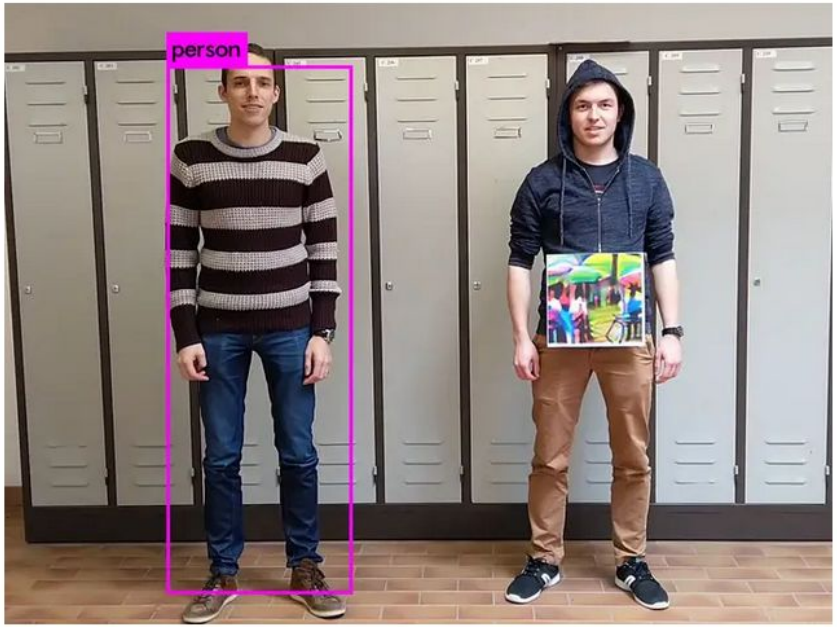
One pixel attack for fooling deep neural networks

Jiawei Su, Danilo Vasconcellos Vargas, Sakurai Kouichi

(Submitted on 24 Oct 2017 (v1), last revised 17 Oct 2019 (this version, v7))



Adversarial Patches



Veículos Autônomos

A Systematic Literature Review about the impact of Artificial Intelligence on Autonomous Vehicle Safety

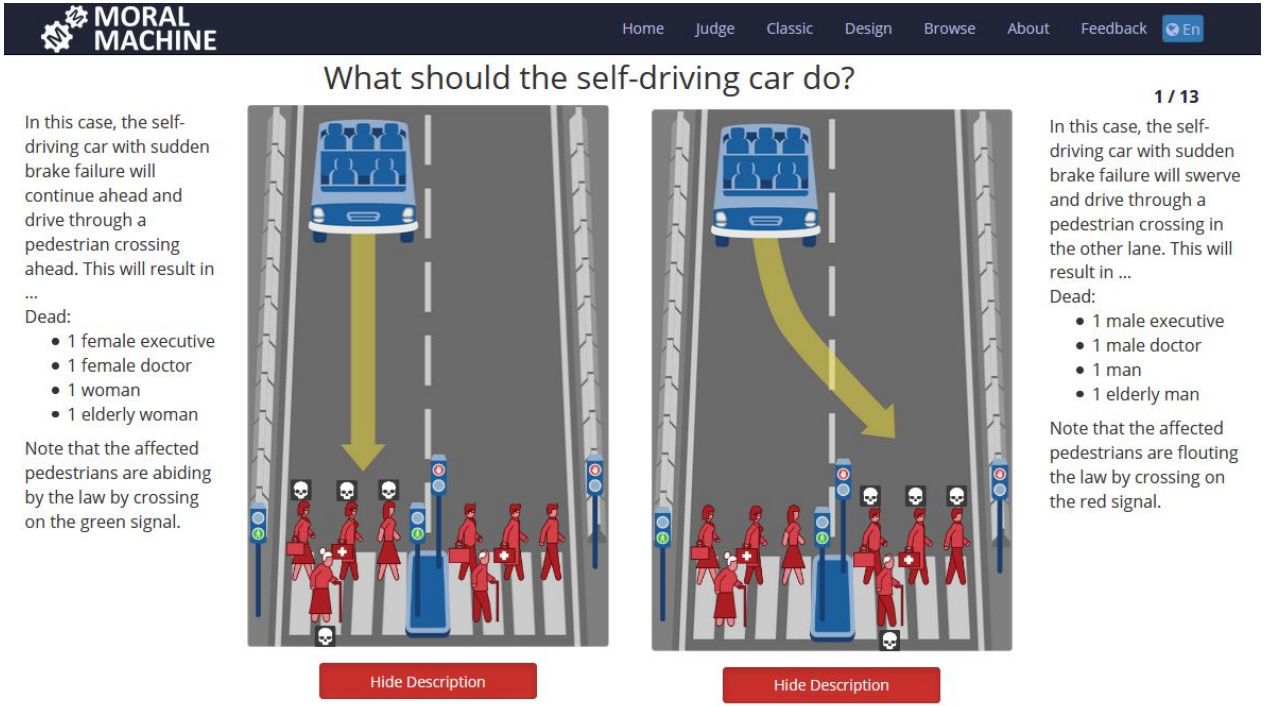
A. M. Nascimento[1], L. F. Vismari[1], C. B. S. T. Molina[1], P.S. Cugnasca[1], J.B. Camargo Jr.[1], J.R. de Almeida Jr.[1], R. Inam[2], E. Fersman[2], M. V. Marquezini[3], and A. Y. Hata[3]

Ethical and Social Aspects of Self-Driving Cars

Tobias Holstein
Mälardalen University
Västerås, Sweden
tobias.holstein@mdh.se

Gordana Dodig-Crnkovic, Patrizio Pelliccione
Chalmers University of Technology | University of Gothenburg
Gothenburg, Sweden
[gordana.dodig-crnkovic,patrizio]@chalmers.se

Podemos confiar em ML?



Automação

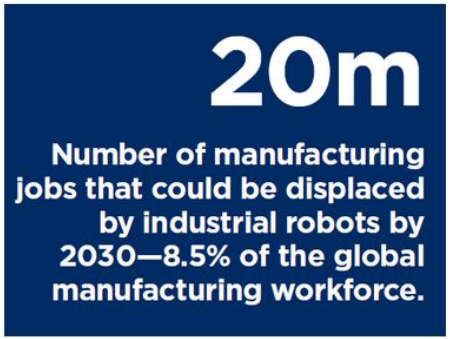
ML e o Mercado de Trabalho

McKinsey Global Institute

Automação ameaça metade dos empregos no país. Saiba as profissões que podem ser afetadas pela tecnologia

Pesquisa estima que 44,5 milhões de brasileiros estão em ocupações em que o homem pode ser substituído por robôs

O Globo



Oxford Economics

Três cenários

- Três formas de encarar o impacto da automação: a boa, a má e a feia.

Cenário bom

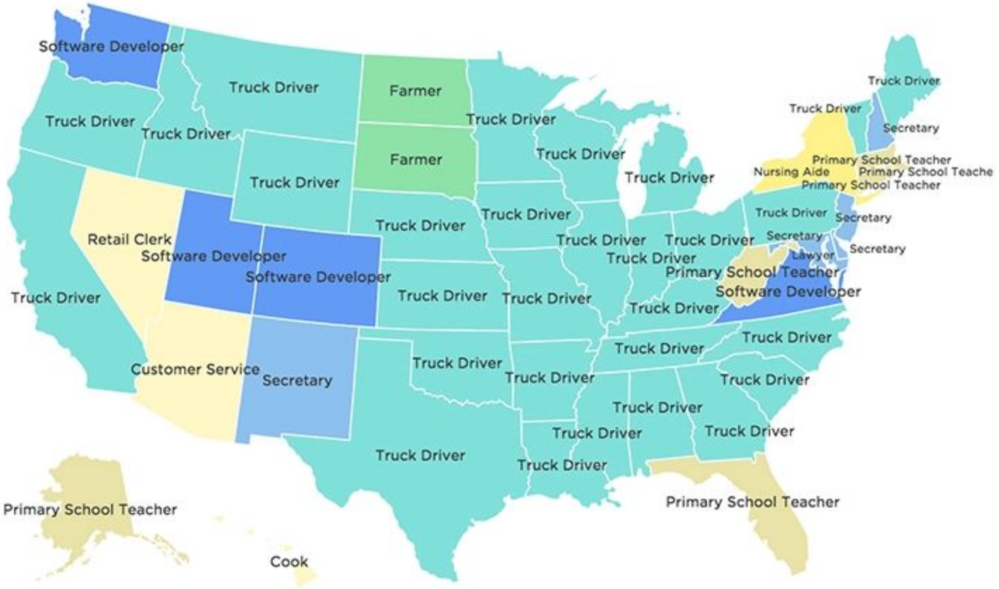
World Economic Forum: automation will displace 75 million jobs but generate 133 million new ones worldwide by 2022.

Gartner: AI-related job creation will reach two million net-new jobs in 2025.

WIRED BUSINESS CULTURE GEAR IDEAS SCIENCE SECURITY TRANSPORTATION

AI May Not Kill Your Job—Just Change It

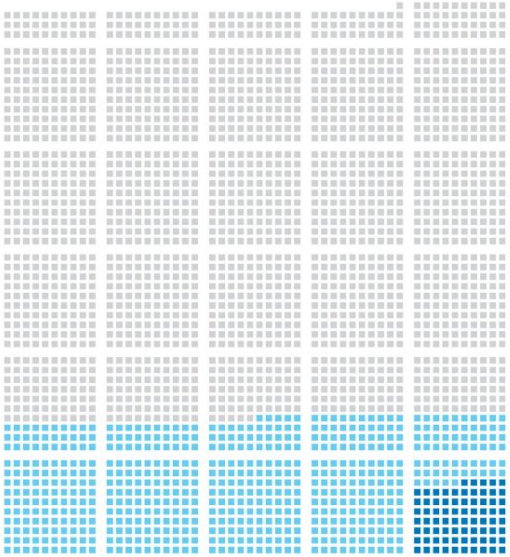
Cenário ruim



Elon Musk: 'A.I. will make jobs kind of pointless'

Published Thu, Aug 29 2019•10:47 AM EDT • Updated Fri, Aug 30 2019•10:54 AM EDT

Cenário feio



Seventy-five million to 375 million may need to switch occupational categories and learn new skills.

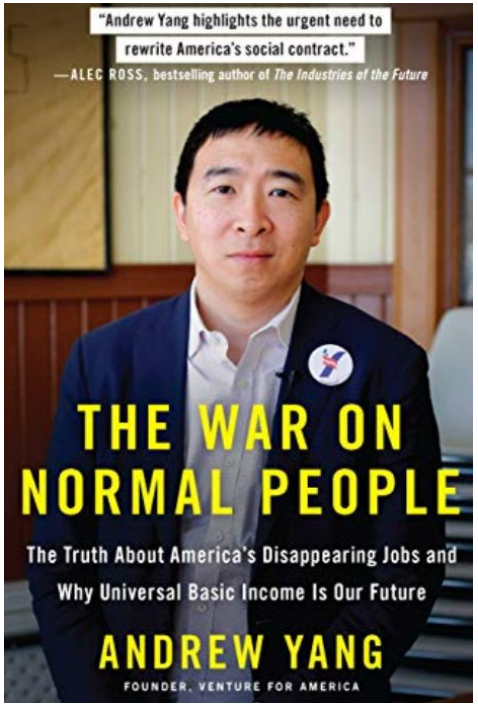
Mckinsey: The future of women at work

Long-established barriers will make it harder for women to make transitions.

Mckinsey: The future of women at work

World Economic Forum warns of AI's potential to worsen global inequality

Renda Básica Universal (UBI)



Perguntas?

Marcelo Prates
marceloorp@gmail.com
linkedin.com/in/marceloprates

Luiz Nonenmacher
ljuniornone@gmail.com
linkedin.com/in/luiz-nonenmacher