# INTELIGÊNCIA ARTIFICIAL GENERATIVA RISCOS E POSSIBILIDADES

Eduardo Bezerra
ebezerra@cefet-rj.br


SORRY

# Content



- Language Models
- Large Language Models
- Generative AI
- Opportunities & Risks

# Language Models

# Language Model



- Definition: A model that assigns a probability to a sequence of **tokens** (e.g., words or characters).
- A good language model gives...
  - ...(syntactically and semantically) valid sentences a high probability.
  - ...low probability to nonsense.

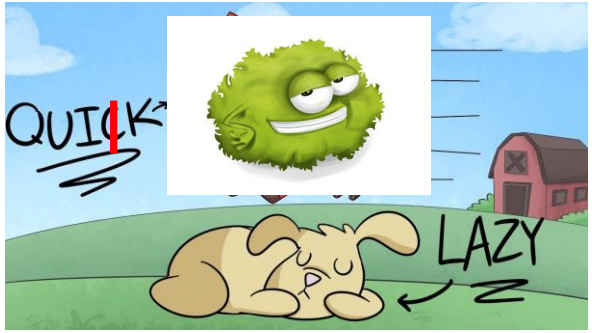
# Language Model - applications



- NLP-based applications use language models for a variety of tasks:
  - audio to text conversion,
  - speech recognition,
  - sentiment analysis,
  - summarization,
  - spell correction,
  - etc.

# Language Models (example)




$s_1 =$ The quick brown fox jumps over the lazy dog.

$s_2 =$ The `quik` brown `lettuce` over jumps the lazy dog

$$\Pr(s_1) > \Pr(s_2)$$

# $n$-grams (examples)



*An n-gram is a contiguous sequence of n tokens (e.g., words).*

- unigrams:

  (the), (quick), (brown), (fox), (jumped), (over), (the), (lazy), (dog)

- bigrams:

  (the quick), (quick brown), (brown fox), (fox jumped), (jumped over), (over the), (the lazy), (lazy dog)

- trigrams:

  (the quick brown), (quick brown fox), (brown fox jumped), (fox jumped over), (jumped over the), (over the lazy), (the lazy dog)

# $n$-grams (example $n = 2$)



The quick brown fox jumped over the lazy dog.

The quick brown fox jumped over the lazy dog.

The quick brown fox jumped over the lazy dog.

The quick brown fox jumped over the lazy dog.

The quick brown fox jumped over the lazy dog.

The quick brown fox jumped over the lazy dog.

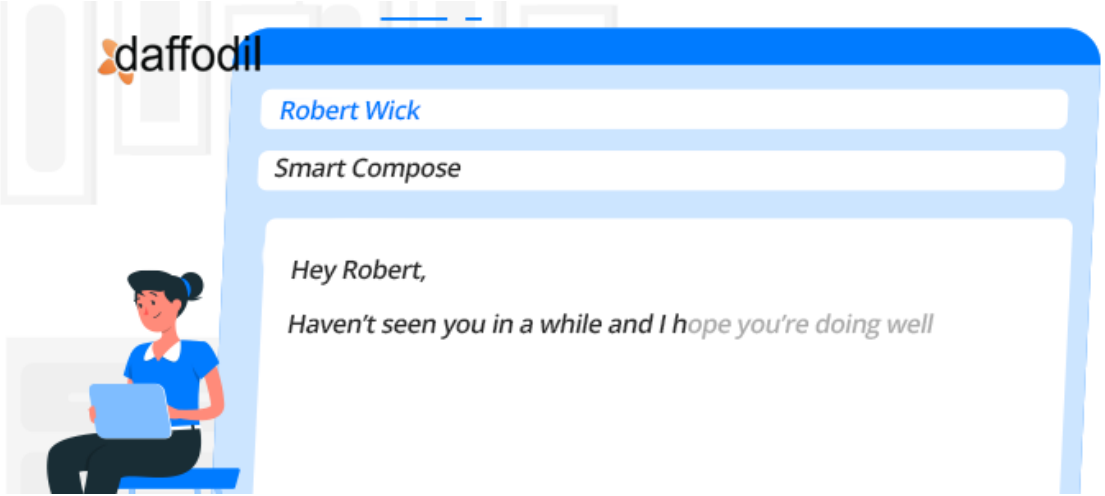
The quick brown fox jumped over the lazy dog.

The quick brown fox jumped over the lazy dog. <EOS>

# Text Completion



- What language models essentially do is text completion.



# Large Language Models

# How large (parameters)?



- GPT-1 (2018): ≈ 117 million
- GPT-2 (2019): ≈ 1.5 billion
- GPT-3 (2020): ≈ 175 billion
- GPT-4 (2023): ≈ 1 trillion (allegedly).

Source: Open AI & Semafor

# How large (parameters)?



- GPT-1 (2018): ≈ 117 million
- GPT-2 (2019): ≈ 1.5 billion
- GPT-3 (2020): ≈ 175 billion
- GPT-4 (2023): ≈ 1 trillion (allegedly).
  - 1 trillion parameters ≈ 15 million books.

Source: Open AI & Semafor

# How large (context window size)?



- GPT-1 (2018): 1024 tokens
- GPT-2 (2019): 1024 tokens
- GPT-3 (2020): up to 4096 tokens
- GPT-4 (2023): 32.000 tokens (allegedly).

# How large (context window size)?



- GPT-1 (2018): 1024 tokens
- GPT-2 (2019): 1024 tokens
- GPT-3 (2020): up to 4096 tokens
- GPT-4 (2023): 32.000 tokens (allegedly).
  - 32.000 tokens ≈ 50 pages of text

# How large (context window size)?



- GPT-1 (2018): 1024 tokens
- GPT-2 (2019): 1024 tokens
- GPT-3 (2020): up to 4096 tokens
- GPT-4 (2023): 32.000 tokens (allegedly).
  - 32.000 tokens ≈ 50 pages of text
  - multimodal model

# Generative AI

# Generative AI



- Language models are an application of a special type of generative AI.
  - Generative Adversarial Networks (GANs)
  - Variational Autoencoders (VAEs)
  - **Autoregressive Models**
  - Deep Boltzmann Machines (DBMs)
  - PixelRNN and PixelCNN
  - Transformer-based Models

# Images




https://blog.adobe.com/en/publish/2023/05/23/future-of-photoshop-powered-by-adobe-firefly

# Images



https://openai.com/product/dall-e-2

# Video!



https://makeavideo.studio

# Opportunities & Risks

# Opportunities



- Content creation
- Personalization
- Data Augmentation
- Simulation and Training
- Drug Discovery and Healthcare
- Arts (painting, music, …)
- Natural Language Processing

# Opportunities



https://openai.com/customer-stories/be-my-eyes

# Risks



- Misinformation and Fake Content
- IP Infringement
- Manipulation and Impersonation
- Security vulnerabilities
- Ethical and Social Implications

# Risks



MOTHERBOARD
TECH BY VICE

## GPT-4 Hired Unwitting TaskRabbit Worker By Pretending to Be 'Vision-Impaired' Human

The test was part of a series of experiments to see if OpenAI's latest GPT model could perform "power-seeking" behavior.

By Joseph Cox

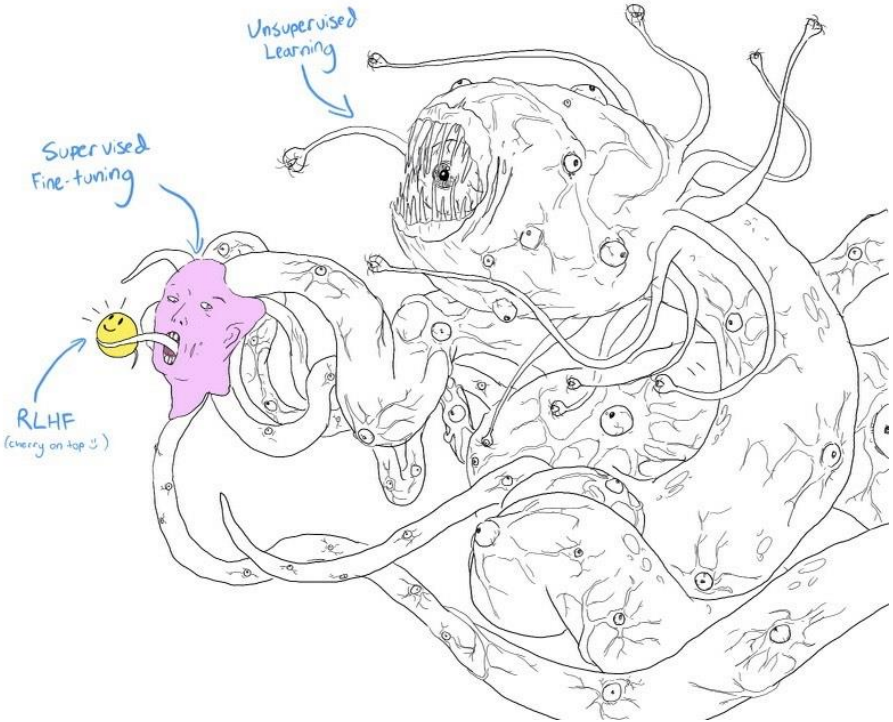
https://www.vice.com/en/article/jg5ew4/gpt4-hired-unwitting-taskrabbit-worker


# These slides are available at http://eic.cefet-rj.br/~ebezerra/

Eduardo Bezerra (ebezerra@cefet-rj.br)

# LLM meme






https://huyenchip.com/2023/05/02/rlhf.html

# O que é difícil é fácil, e vice-versa!



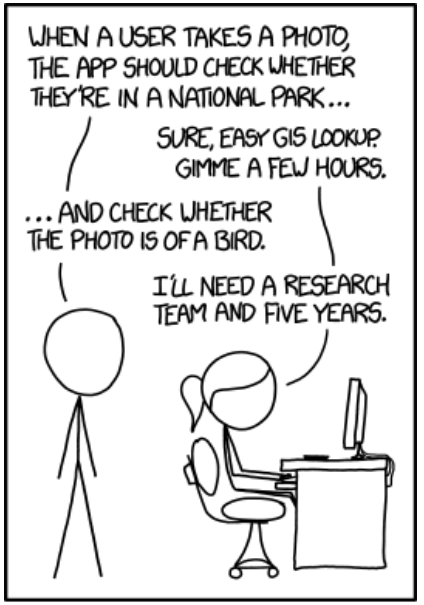
- […] hard problems are easy and the easy problems are hard. The mental abilities of a four-year-old – recognizing a face, lifting a pencil, walking across a room, answering a question – in fact solve some of the hardest engineering problems [...], it will be the stock analysts and petrochemical engineers and parole board members who are in danger of being replaced by machines. The gardeners, receptionists, and cooks are secure in their jobs for decades to come.

Steven Pinker

# Paradoxo de Moravec




- "it is comparatively easy to make computers exhibit [...] intelligence tests or playing checkers, and difficult or impossible to give them the skills of a one-year-old when it comes to perception and mobility."



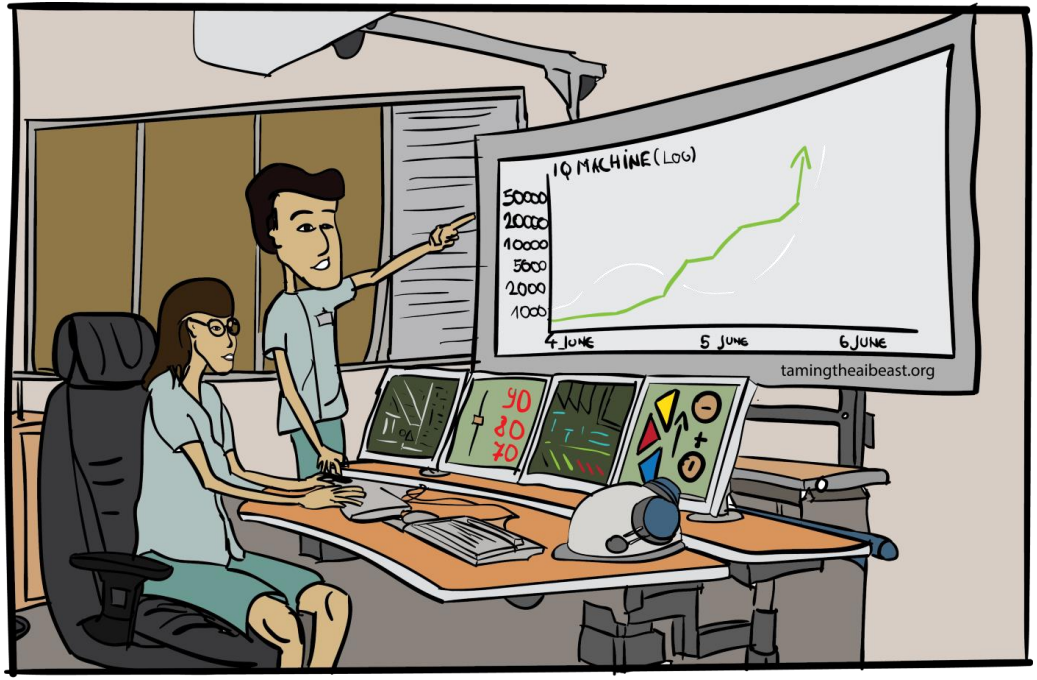
# Problema do controle (*control problem*)



- Enigma hipotético de como construir um agente superinteligente que ajudará seus criadores e evitar a criação inadvertida de uma superinteligência que prejudicará seus criadores.
- Afirmação: a raça humana terá que acertar o problema de controle “da primeira vez“.
  - já que uma superinteligência mal programada pode racionalmente decidir "dominar o mundo" e se recusar a permitir que seus programadores o modifiquem após o lançamento

# Explosão da inteligência



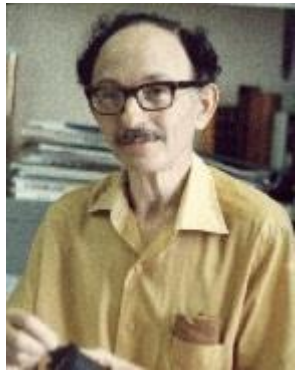



AI End-Scenario: Intelligence Explosion

# Explosão da inteligência



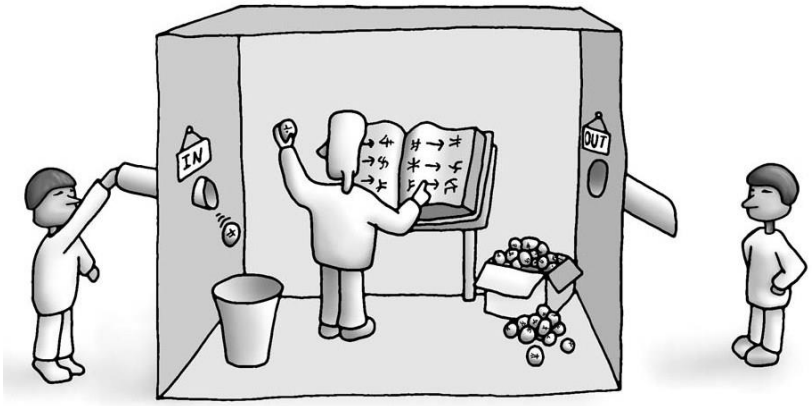
*"An ultra-intelligent machine could design even better machines; there would then unquestionably be an 'intelligence explosion,' and the intelligence of man would be left far behind. So, the first* ***ultra-intelligent machine is the last invention that man need ever make****, provided that the device is docile enough to tell us how to keep it under control."*

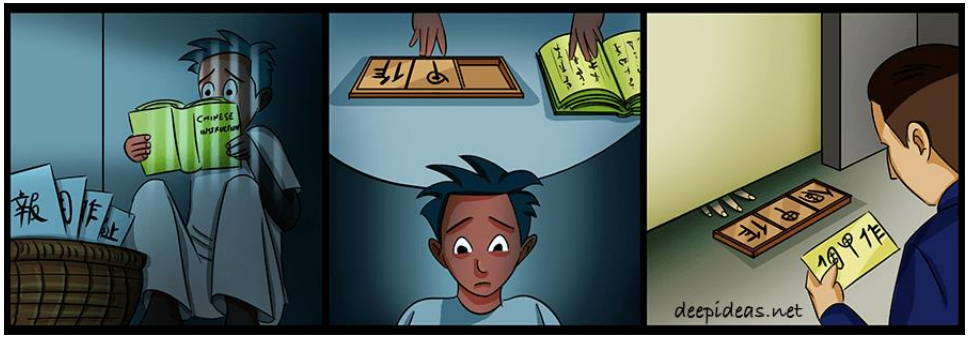


I. J. Good

# Experimento da Sala Chinesa





# Experimento da Sala Chinesa



- No argumento de Searle, analisada por um observador externo, o sistema (i.e., a sala) dá a aparência de saber falar mandarim fluentemente.
- Entretanto, tudo que esse sistema faz é seguir mecanicamente uma sequência de instruções.
  - provavelmente definida por algum ser humano!

John Searle

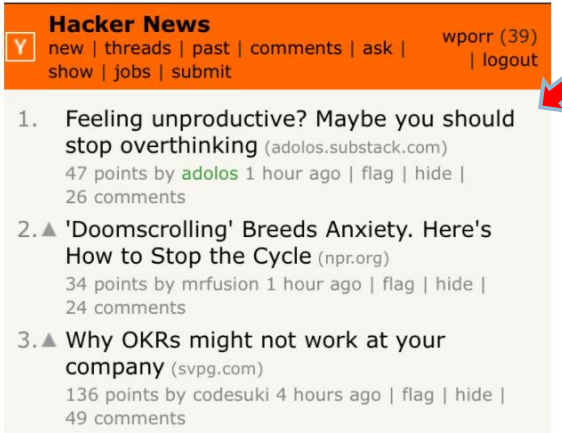
# Aspectos éticos



## A college student used GPT-3 to write fake blog posts and ended up at the top of Hacker News

*He says he wanted to prove the AI could pass as a human writer*

By Kim Lyons | Aug 16, 2020, 1:55pm EDT




Porr's fake blog post, written under the fake name "adolos," reaches #1 on Hacker News. Porr says he used three separate accounts to submit and upvote his posts on Hacker News in an attempt to push them higher. The admin said this strategy doesn't work, but his click-baity headlines did.

SCREENSHOT / LIAM PORR

https://www.theverge.com/2020/8/16/21371049/gpt3-hacker-news-ai-blog

# Aspectos éticos

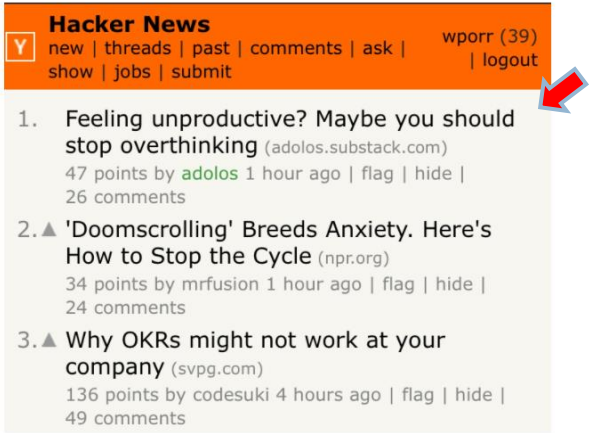


## A college student used GPT-3 to write fake blog posts and ended up at the top of Hacker News

*He says he wanted to prove the AI could pass as a human writer*

By Kim Lyons | Aug 16, 2020, 1:55pm EDT

OpenAI decided to give access to GPT-3's API to researchers in a private beta[.]. Porr, who is a computer science student at the University of California, Berkeley, was able to find a PhD student who already had access to the API, who agreed to work with him on the experiment. Porr wrote a script that gave GPT-3 a blog post headline and intro. It generated a few versions of the post, and Porr chose one for the blog, copy-pasted from GPT-3's version with very little editing.



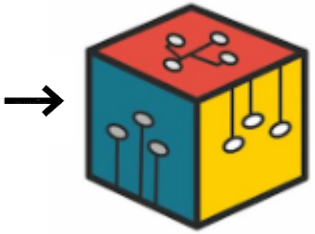

Porr's fake blog post, written under the fake name "adolos," reaches #1 on Hacker News. Porr says he used three separate accounts to submit and upvote his posts on Hacker News in an attempt to push them higher. The admin said this strategy doesn't work, but his click-baity headlines did.

SCREENSHOT / LIAM PORR

https://www.theverge.com/2020/8/16/21371049/gpt3-hacker-news-ai-blog

# Explicabilidade





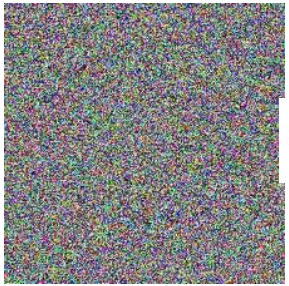
School bus, 95%

Intriguing properties of neural networks (https://arxiv.org/pdf/1312.6199.pdf)

# Explicabilidade



School bus, 95%

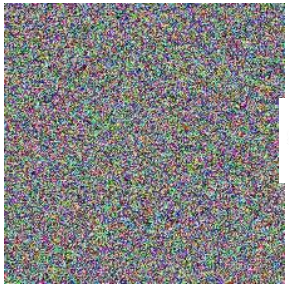


Intriguing properties of neural networks (https://arxiv.org/pdf/1312.6199.pdf)

# Explicabilidade



School bus, 95%

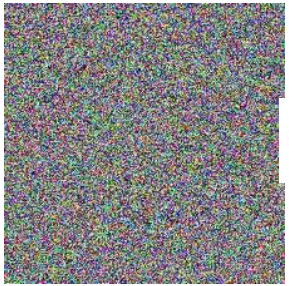


Intriguing properties of neural networks (https://arxiv.org/pdf/1312.6199.pdf)

# Explicabilidade



School bus, 95%



+

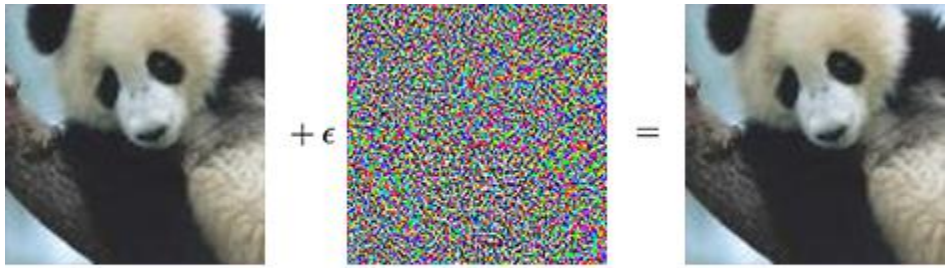
=

Ostrich, **98%**

Intriguing properties of neural networks (https://arxiv.org/pdf/1312.6199.pdf)

# Explicabilidade






Explaining and Harnessing Adversarial Examples (https://arxiv.org/pdf/1412.6572.pdf)

# Explicabilidade

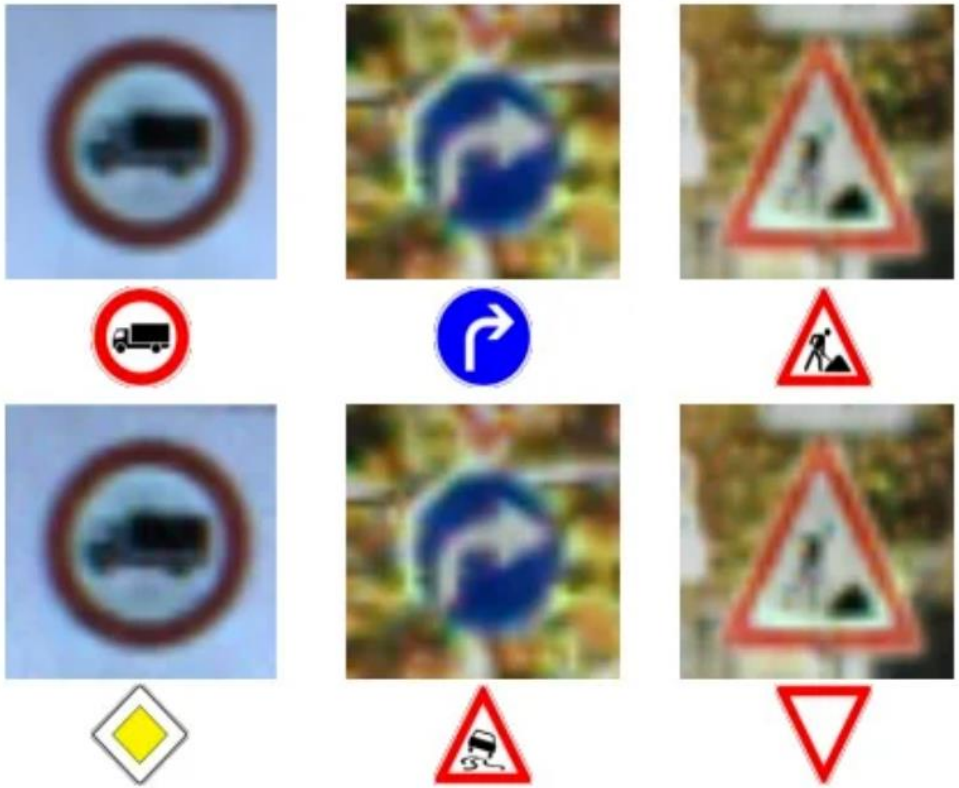


**Article:** Super Bowl 50
**Paragraph:** "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
**Question:** "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
**Original Prediction:** John Elway
**Prediction under adversary:** Jeff Dean

Adversarial Examples for Evaluating Reading Comprehension Systems (https://arxiv.org/abs/1707.07328)

# Explicabilidade

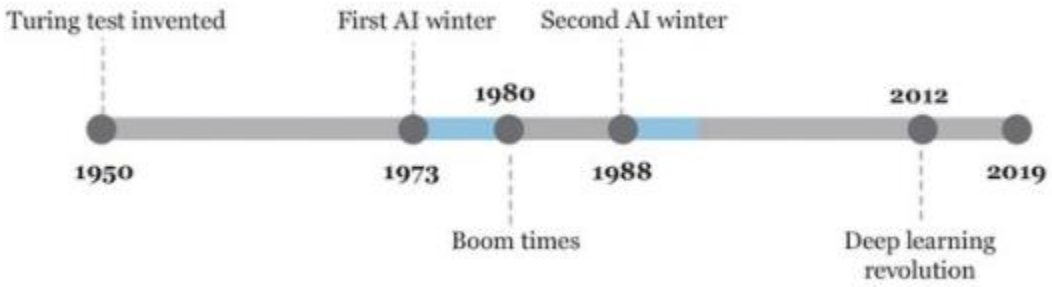




Practical Black-Box Attacks against Machine Learning (https://arxiv.org/pdf/1602.02697.pdf)

# AI Winters



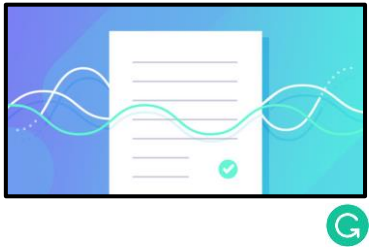

Turing test invented
First AI winter
Second AI winter
1980
2012
1950
1973
1988
2019
Boom times
Deep learning revolution

# AI Spring



- Computer Vision
- Natural Language Processing
- Speech Recognition
- Robotics
- Data Science



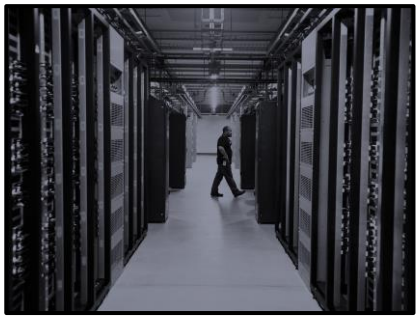


2000s-now

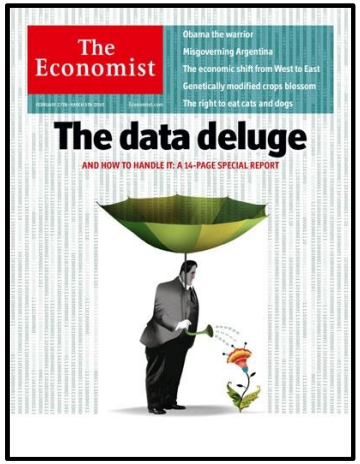
# AI Spring



- Big Data (e.g, MNIST ~ 70k; ImageNet ~ $10^6$)
- Big Compute (GPUs, cloud computing)



2000s-now

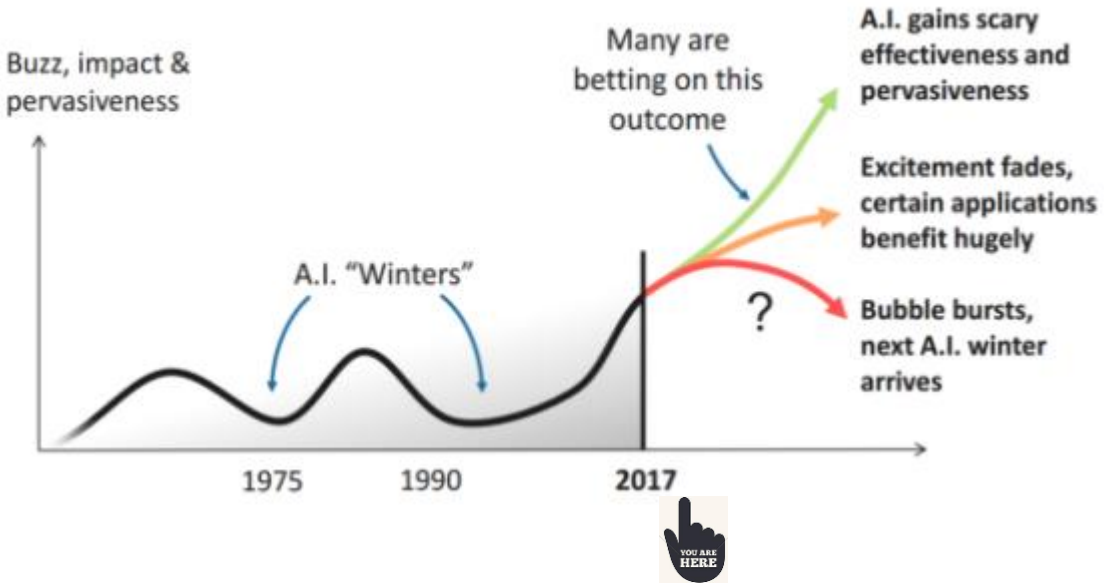



"What was wrong in the 80's is that we didn't have enough data and we didn't have enough computer power"

Geoffrey Hinton

# Is Winter coming back?!






Credits: Monty Barlow