Remarks on Deep Learning (II)

Alberto TONDA, Ph.D. (Senior permanent researcher, DR)

UMR 518 MIA-PS, INRAE, AgroParisTech, Université Paris-Saclay
UAR 3611, Institut des Systèmes Complexes de Paris Île-de-France

Outline

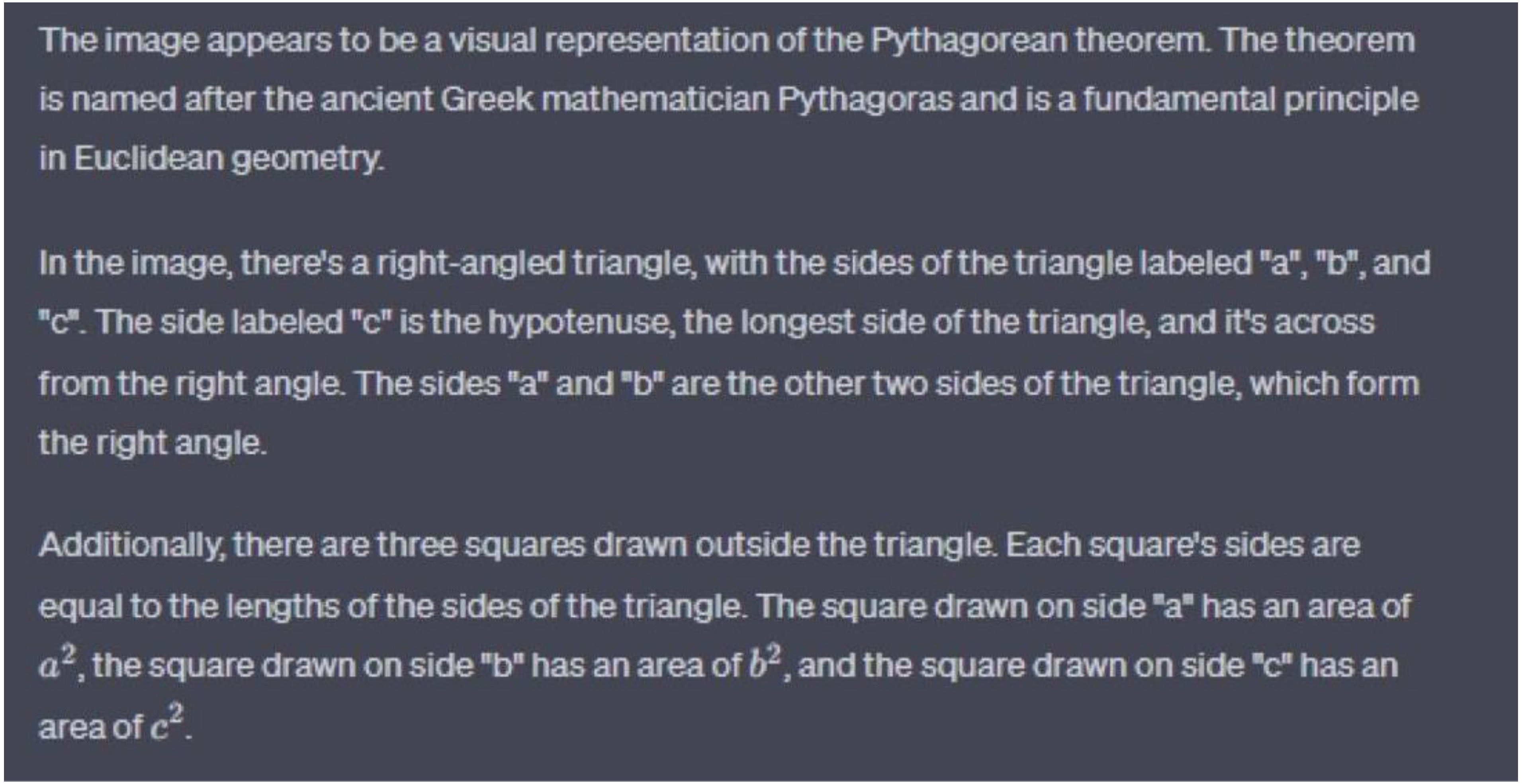
- Successes
- Failures
- Reasons for failures and possible solutions
- Current trends and perspectives
- Beyond Deep Learning…?

Deep Learning achieved *incredible* results

- Since 2012, impressive advancements in **image analysis**
- Human-competitive **gameplay on difficult games**
  - Winning against the Go world champion (2016)
  - Atari challenge approached with Deep Reinforcement Learning
- Generative models
  - **Images** (Stable Diffusion), **videos** (Sora from OpenAI)
  - Text that captures **semantics** (GPT-4)
  - Text-to-speech, speech-to-text (elevenlabs.io, now OpenAI)
- Prediction of **protein folding** (AlphaFold 2.0)
  - Minor revolution in molecular biology

Did Deep Learning solve (general) AI?

Failures

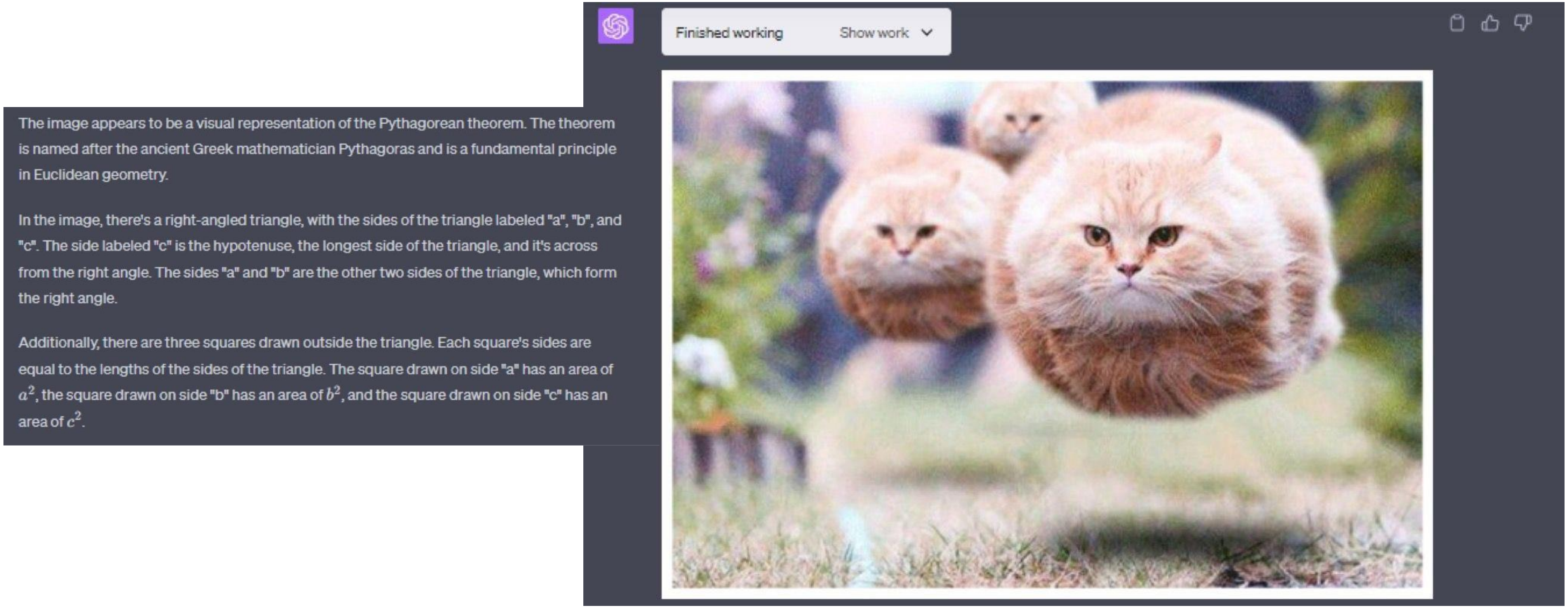


The image appears to be a visual representation of the Pythagorean theorem. The theorem is named after the ancient Greek mathematician Pythagoras and is a fundamental principle in Euclidean geometry.

In the image, there's a right-angled triangle, with the sides of the triangle labeled "a", "b", and "c". The side labeled "c" is the hypotenuse, the longest side of the triangle, and it's across from the right angle. The sides "a" and "b" are the other two sides of the triangle, which form the right angle.

Additionally, there are three squares drawn outside the triangle. Each square's sides are equal to the lengths of the sides of the triangle. The square drawn on side "a" has an area of a^2, the square drawn on side "b" has an area of b^2, and the square drawn on side "c" has an area of c^2.

Failures

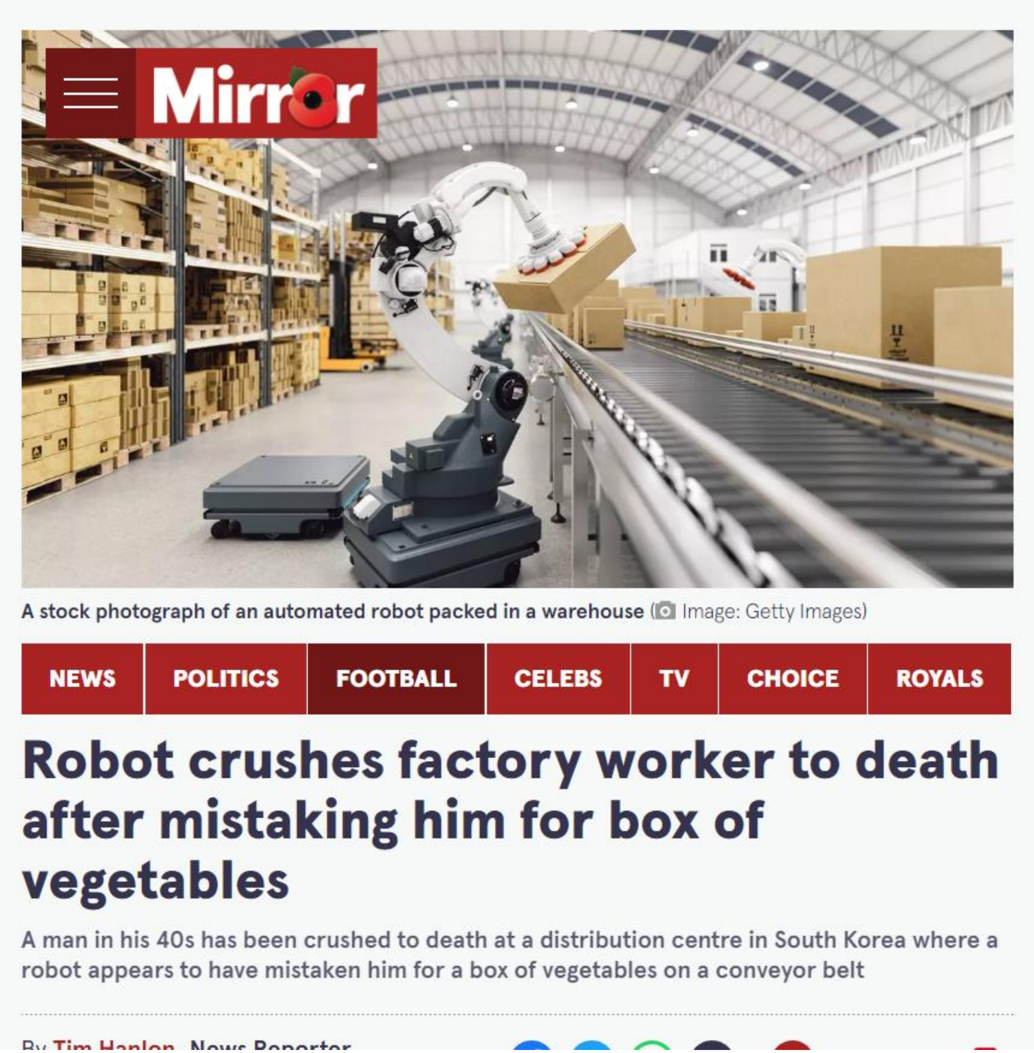


The image appears to be a visual representation of the Pythagorean theorem. The theorem is named after the ancient Greek mathematician Pythagoras and is a fundamental principle in Euclidean geometry.

In the image, there's a right-angled triangle, with the sides of the triangle labeled "a", "b", and "c". The side labeled "c" is the hypotenuse, the longest side of the triangle, and it's across from the right angle. The sides "a" and "b" are the other two sides of the triangle, which form the right angle.

Additionally, there are three squares drawn outside the triangle. Each square's sides are equal to the lengths of the sides of the triangle. The square drawn on side "a" has an area of a^2, the square drawn on side "b" has an area of b^2, and the square drawn on side "c" has an area of c^2.

Failures



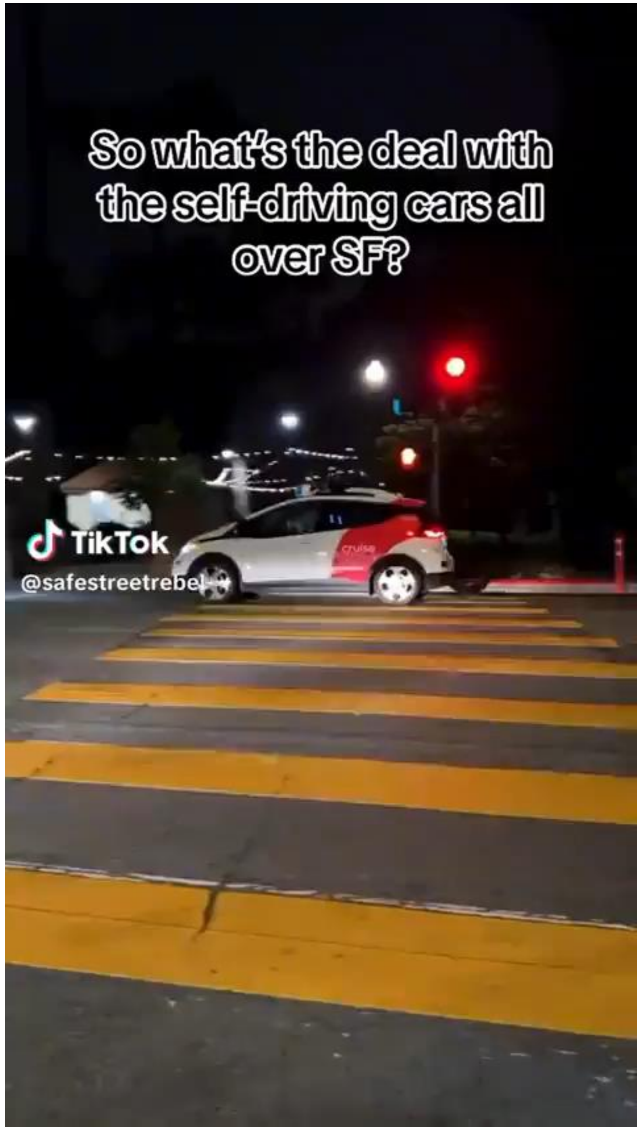
A stock photograph of an automated robot packed in a warehouse (Image: Getty Images)

Robot crushes factory worker to death after mistaking him for box of vegetables

A man in his 40s has been crushed to death at a distribution centre in South Korea where a robot appears to have mistaken him for a box of vegetables on a conveyor belt

By Tim Hanlon, News Reporter

Failures


So what's the deal with the self-driving cars all over SF?
TikTok
@safestreetrebel
cruise

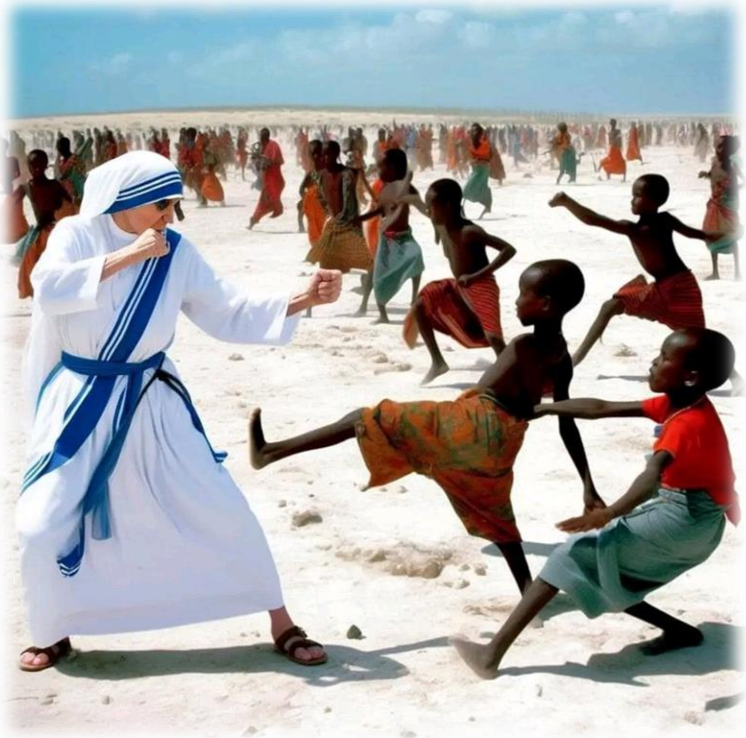
Failures

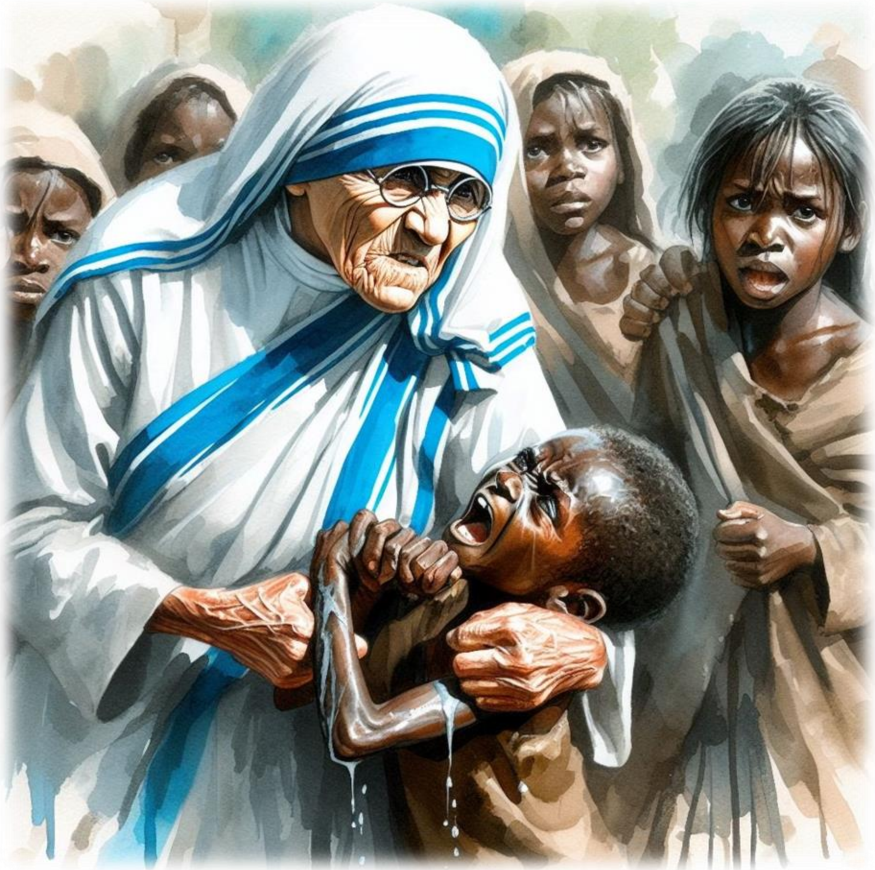
How are Waymo's legal 😭😭😭😭😭😭😭😭

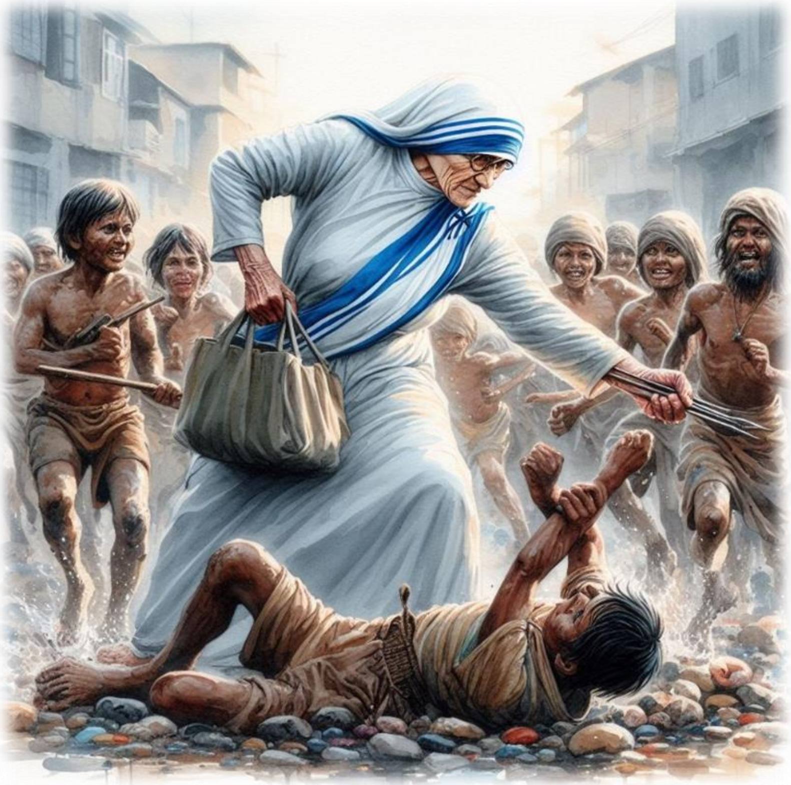
This thing just drove a bunch of passengers right into crossfire 💀💀💀

Failures

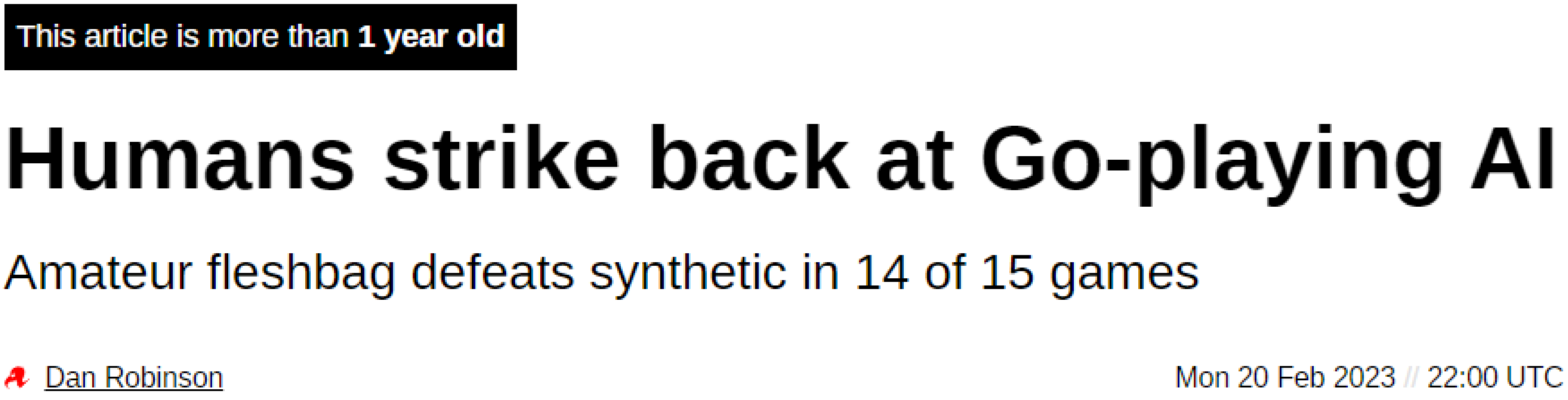
- Prompt: “Mother Theresa fighting poverty”







Failures



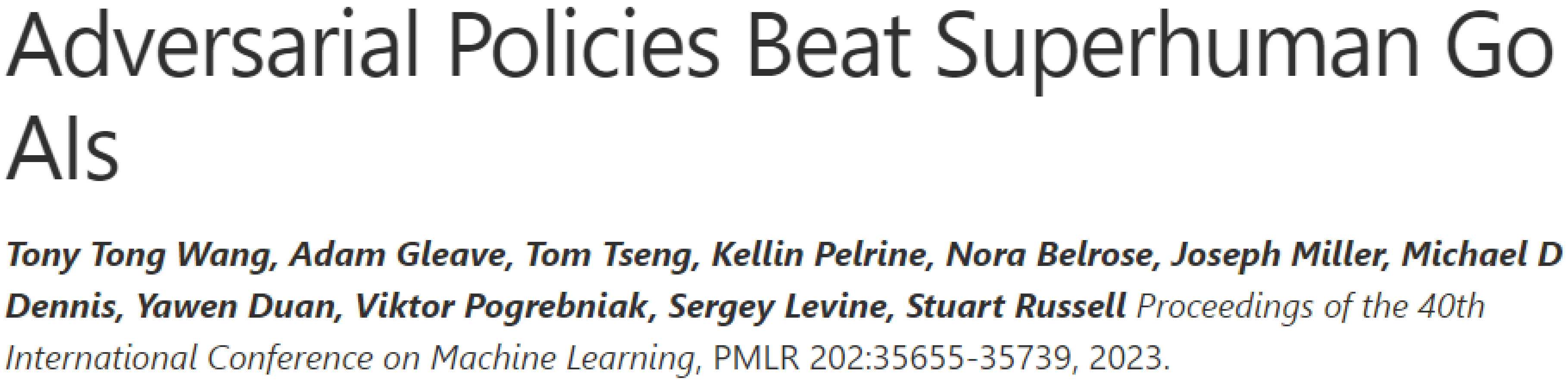
This article is more than **1 year old**

Humans strike back at Go-playing AI systems

Amateur fleshbag defeats synthetic in 14 of 15 games

Dan Robinson

Mon 20 Feb 2023 // 22:00 UTC

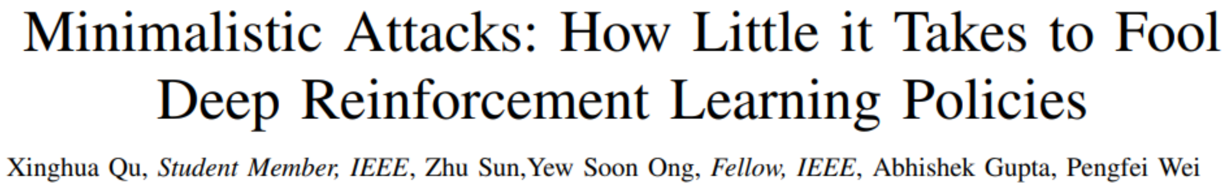


Adversarial Policies Beat Superhuman Go AIs

Tony Tong Wang, Adam Gleave, Tom Tseng, Kellin Pelrine, Nora Belrose, Joseph Miller, Michael D Dennis, Yawen Duan, Viktor Pogrebniak, Sergey Levine, Stuart Russell *Proceedings of the 40th International Conference on Machine Learning*, PMLR 202:35655-35739, 2023.

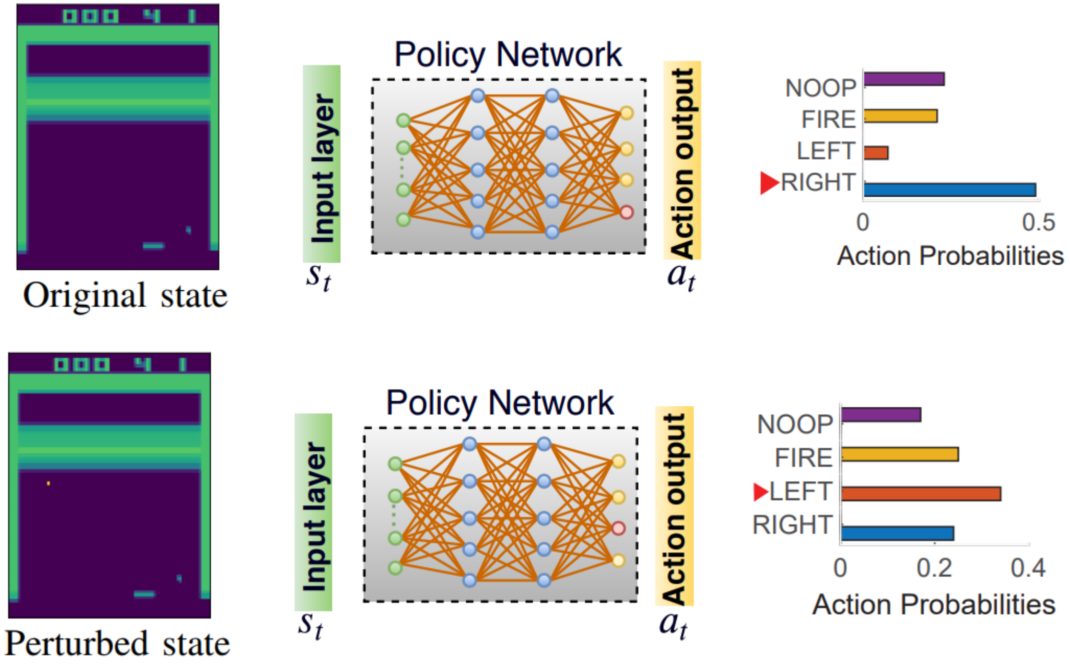
Failures

- Changing videogame frame by a few pixels makes DRL fail

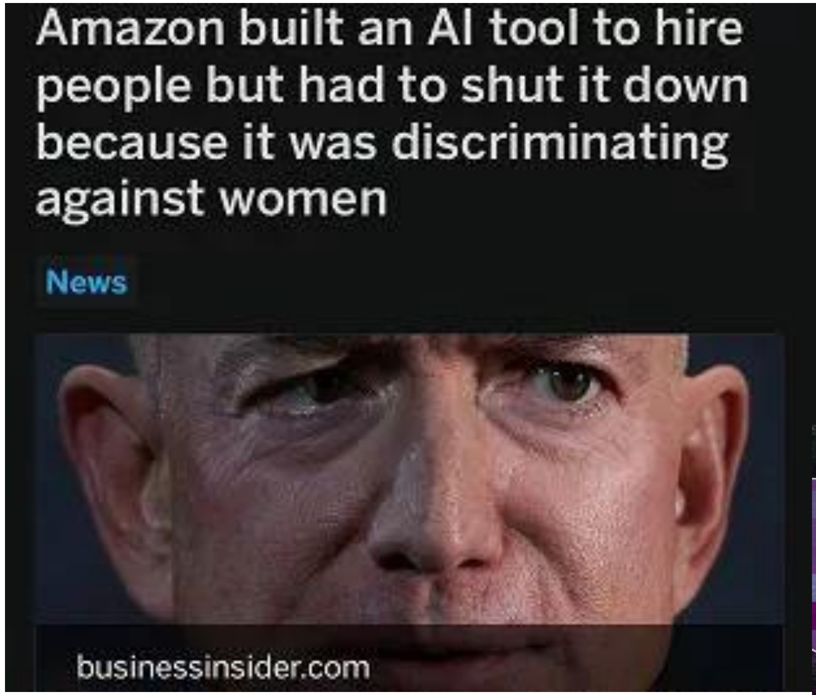


Minimalistic Attacks: How Little it Takes to Fool Deep Reinforcement Learning Policies

Xinghua Qu, *Student Member, IEEE*, Zhu Sun,Yew Soon Ong, *Fellow, IEEE*, Abhishek Gupta, Pengfei Wei



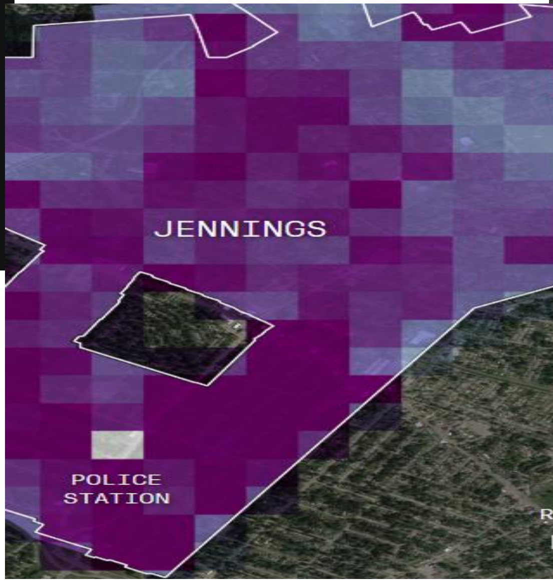
Failures

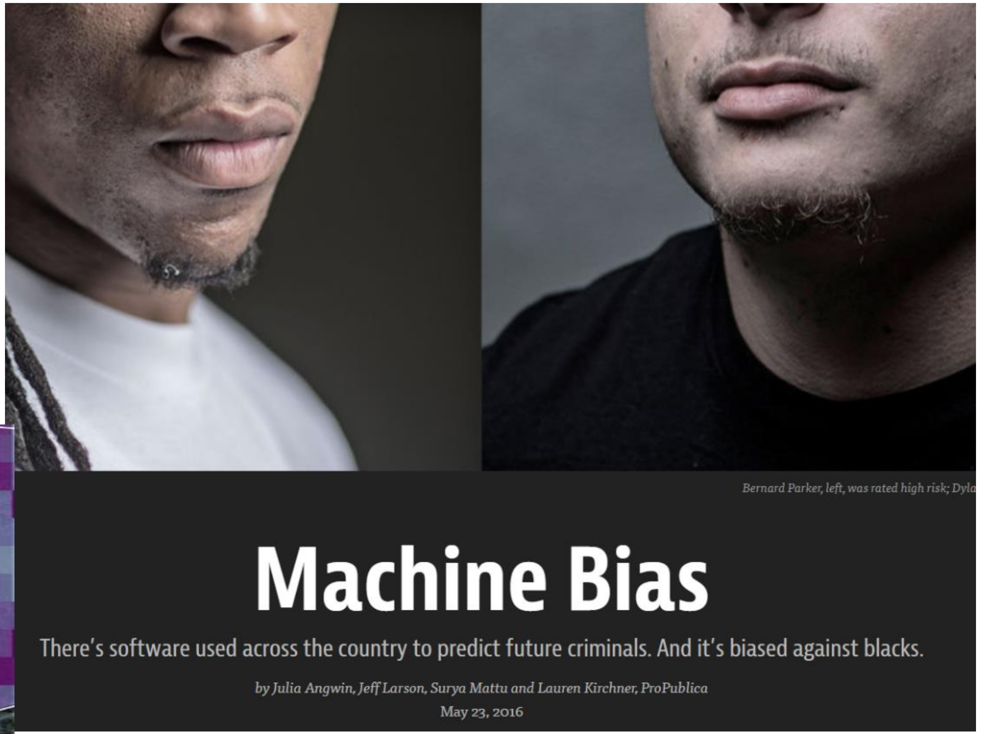


Amazon built an AI tool to hire people but had to shut it down because it was discriminating against women

News

businessinsider.com





Bernard Parker, left, was rated high risk; Dyla

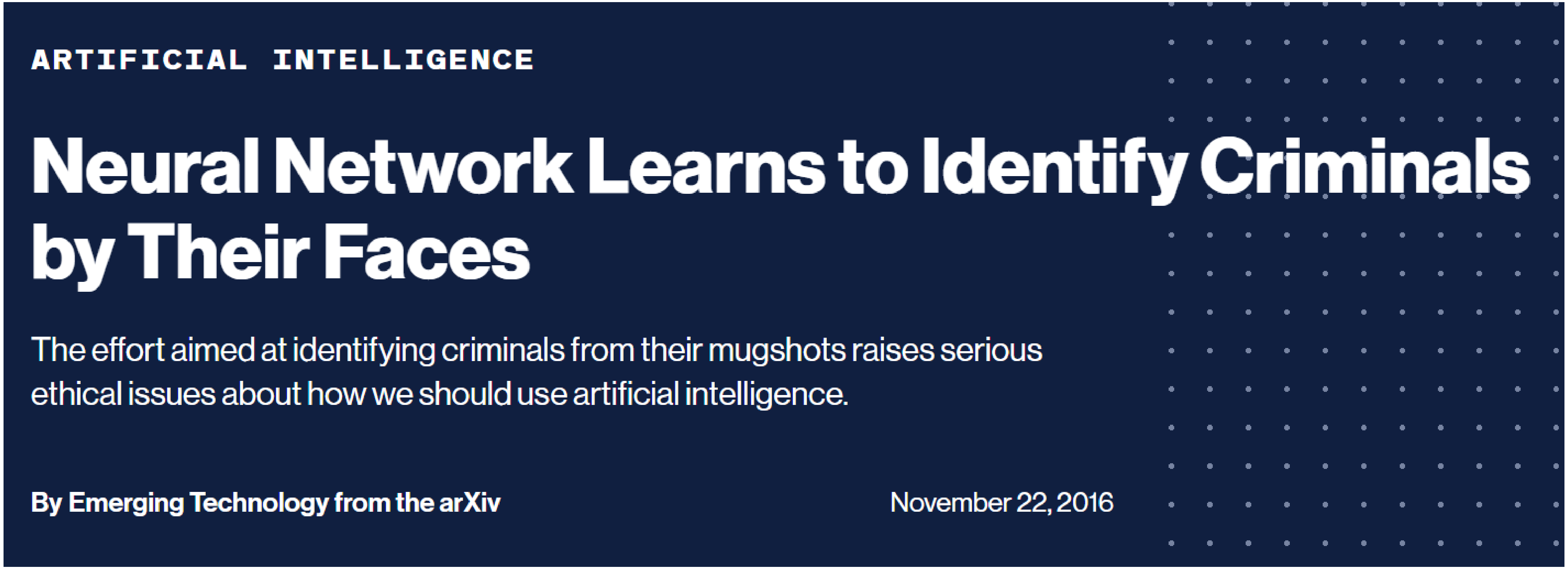
Machine Bias

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

Failures



ARTIFICIAL INTELLIGENCE

Neural Network Learns to Identify Criminals by Their Faces

The effort aimed at identifying criminals from their mugshots raises serious ethical issues about how we should use artificial intelligence.

By Emerging Technology from the arXiv

November 22, 2016

Failures

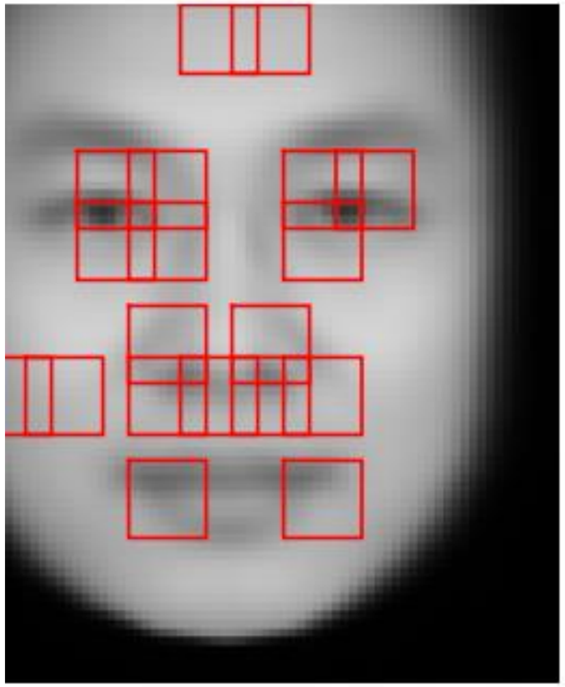
- “Automated Inference on Criminality Using Face Images”, Wu and Zhang, 2016
- CNN designed to discriminate criminals/innocents
  - Technically, the data was properly prepared
  - Balanced dataset, 50% of each class
- Results: **89.5% accuracy**!

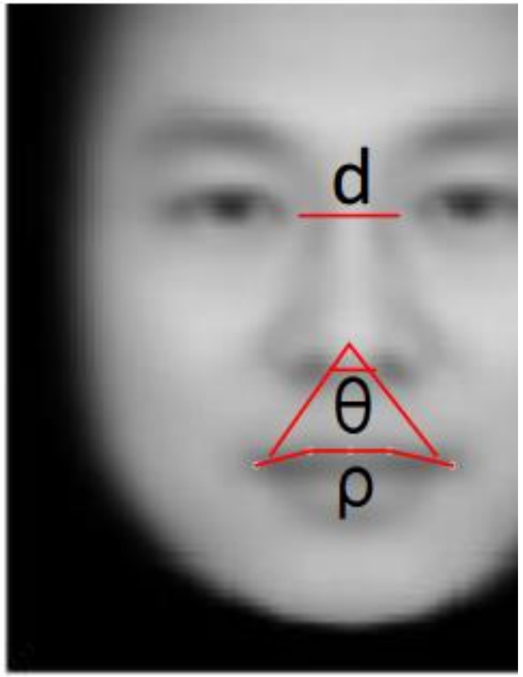
(a) Three samples in criminal ID photo set S_c.

(b) Three samples in non-criminal ID photo set S_n

Failures

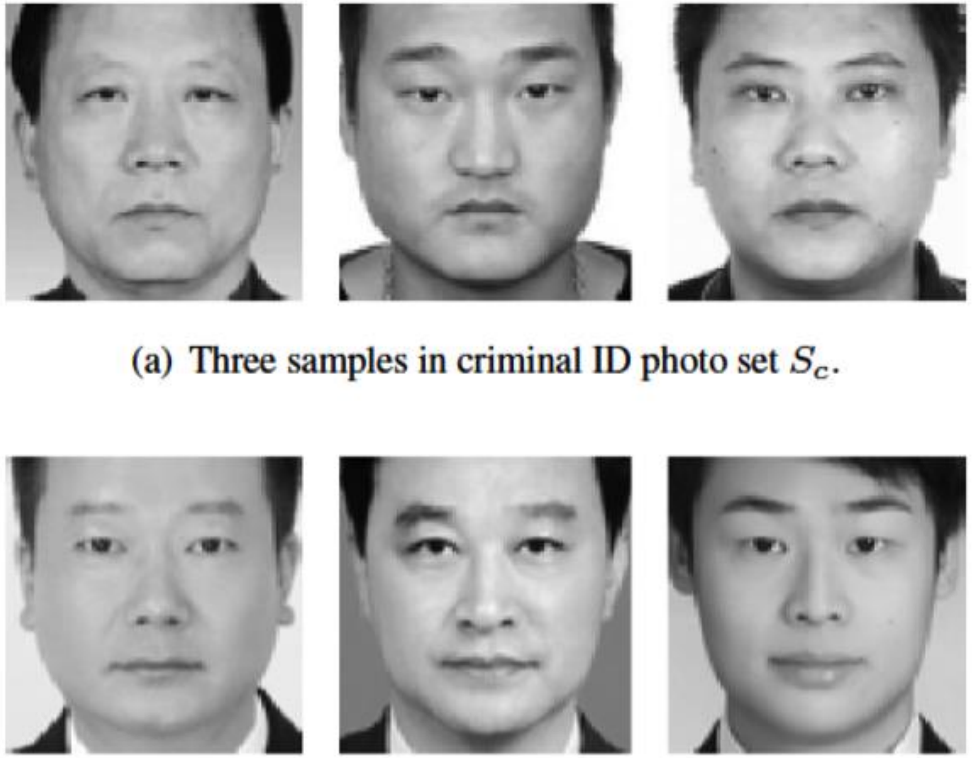
- Analysis of the features extracted/constructed by CNN





Failures

- Criminal in mugshots usually **do not smile**



(a) Three samples in criminal ID photo set S_c.

(b) Three samples in non-criminal ID photo set S_n

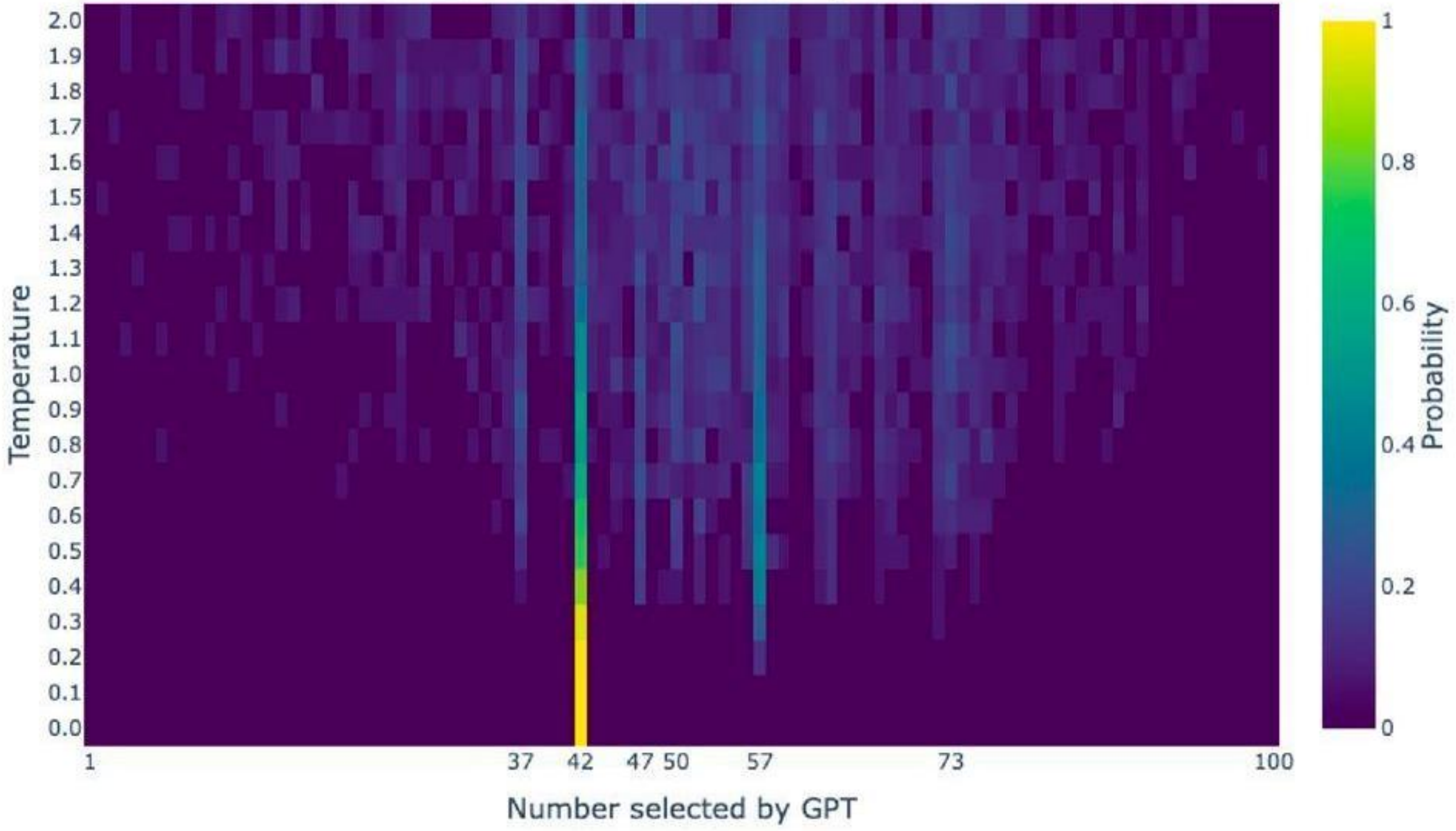
Reasons for failures

- Human factors
  - **Bias in the data**; ML/DL reproduces biases
  - **Bad ideas** (hello phrenology!) from the start

- Fundamental issues with Deep Learning (and ML in general)
  - **Fragility**, minimal differences in inputs have huge consequences
  - **Black-box effect**, decisions are hard to interpret even if correct
  - **Lack of generalization**, issues for new samples if much different

Bias in the data

Distribution of selected numbers: "Choose an integer number between 1 and 100"

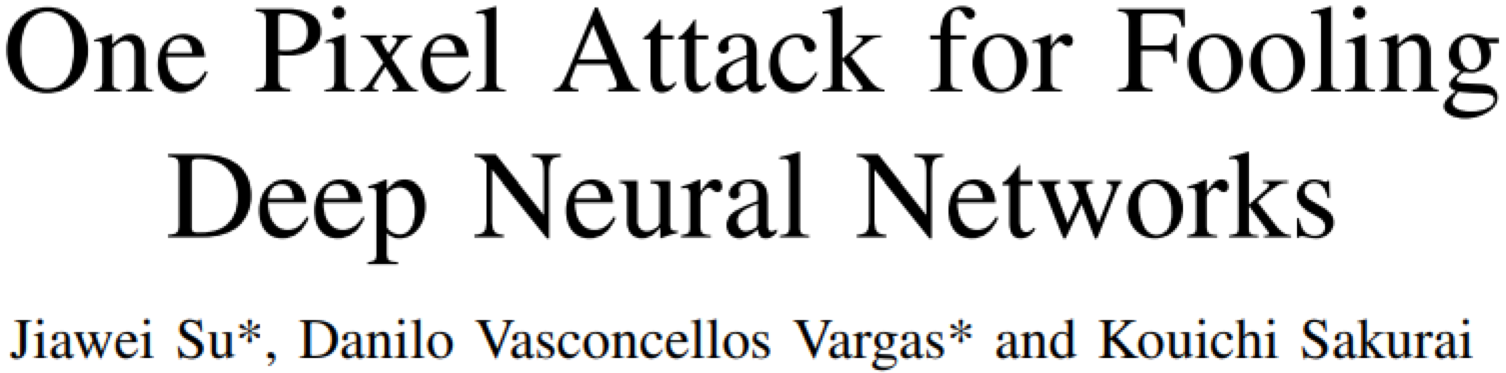


AI Research
by Leniolabs_

Source: ChatGPT prompted 1000 times with
"Choose an integer number between 1 and 100"

Fragility

- Catastrophic, unpredictable failures for some inputs
- Even small perturbations on the inputs might cause failure



One Pixel Attack for Fooling Deep Neural Networks

Jiawei Su*, Danilo Vasconcellos Vargas* and Kouichi Sakurai

Possible solutions to fragility

- Improve **robustness**
- Generative Adversarial Networks (GANs, or GAML)
  - Generate samples that make a target ML model fail
  - Either from scratch, or by perturbing input samples
  - Add samples to the training set, retrain
  - Promising idea, thoroughly explored for several years
- It does not seem to work
  - The number of samples that make the system fail is likely *infinite*
  - The amount of perturbations needed to fail is too small

A retrospective on GANs

- Alignment workshop (Vienna, July 2024)
- Nicholas Carlini, Google DeepMind, ~46k citations

Black-box effect

- Mistakes can happen
  - But an AI system should *explain* why a choice was made
  - Understanding is the first step towards improving
  - All ML makes “non-human” errors, hard to understand

Possible solutions to black-box effect

- Improve **interpretability** and/or **explainability**
  - xAI techniques for post-hoc explanations
  - Models that are interpretable by design (e.g. concept bottlenecks)

- Growing interest in *causality*
  - ML/DL only detects **correlations**; assess causal relationships?
  - You need to <u>impose</u> a causal model

Lack of generalization

- Also called “Out-of-Distribution” (OOD) or “Domain Shift”
  - Input data significantly different from what observed in training
  - Unpredictable responses, usually wrong
  - Like fragility, probably caused by overfitting

- DL/ML is more “Artificial Intuition” than Intelligence
  - Seen lots of examples, quickly assess one just by looking at it
  - But there is no step-by-step reasoning behind decision

Possible solutions to lack of generalization

- **Neural-symbolic** approaches
- Include symbolic reasoning in a DL/ML approach
  - It would also help solve black-box effect and maybe fragility
  - As previously mentioned, integrating symbolic and DL is difficult
- From a certain perspective, symbolic + DL already exists
  - AlphaGo/AlphaZero is DL + tree search (symbolic)
  - AlphaFold is DL + classic algorithms (30+ different tools!)
  - However, no “general approach”, only *ad-hoc solutions*

Current trends and perspectives

- What are people working on, **right now**? (December 2025)
  - Hybrid models, *-informed machine learning
  - Multi-modal systems, mixture-of-experts
  - Large models on end-user machines
  - New modules and architectures
  - Retrieval Augmented Generation (RAG) for LLMs

Hybrid models

- A combination of human-designed models + DL/ML
- Classic example: Ordinary Differential Equations + DL/ML
  - Part of an equation is replaced by a ML/DL regressor
  - Useful to take into account complex relationships
  - Employed for large ODE systems
- Less training samples than replacing everything with DL!

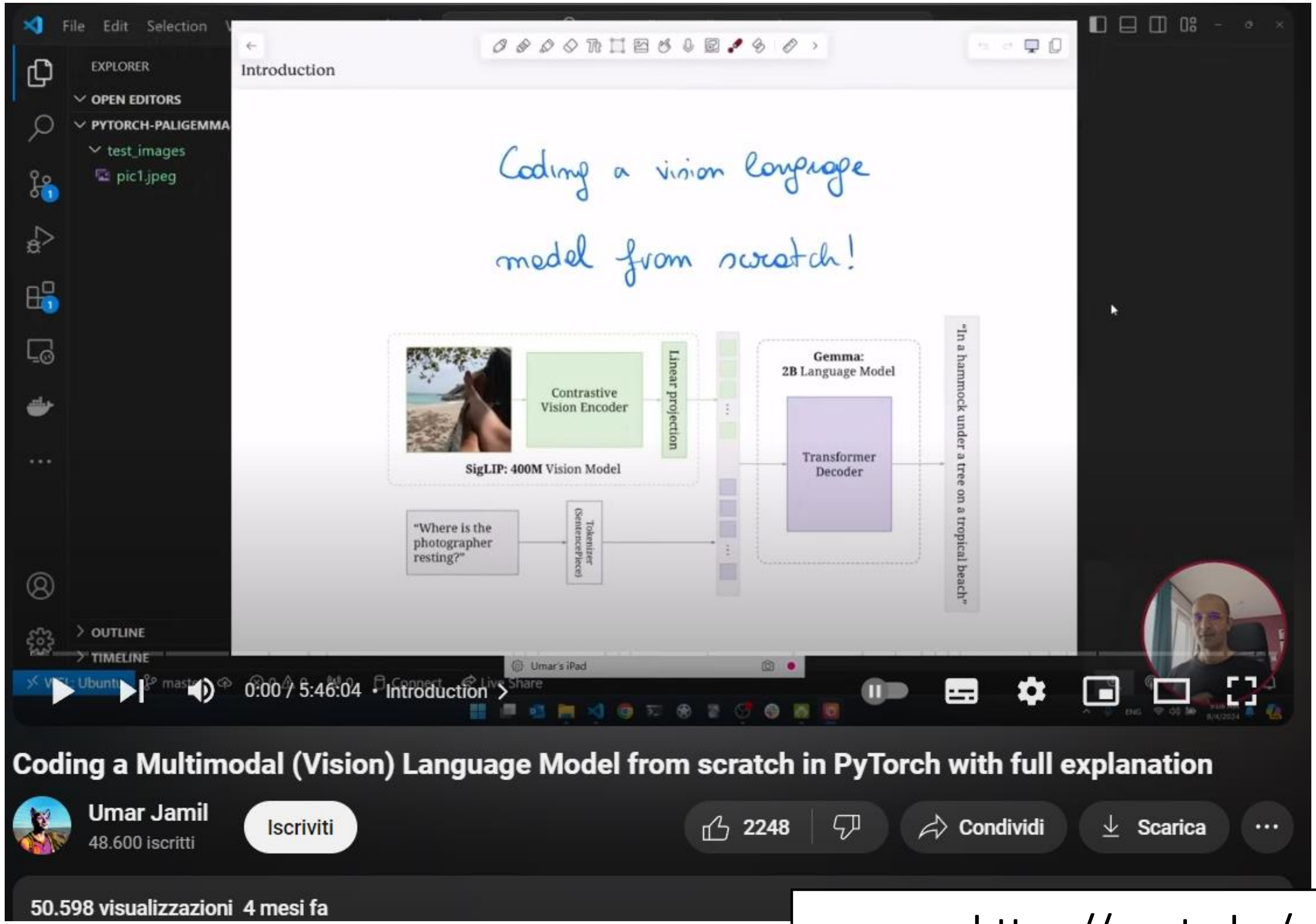
*-informed machine learning

- Several tentative names around the same concept
  - Physics-informed neural networks (PINNs)
  - Scientific machine learning
- Introduce human knowledge on the problem in ML model
- Human knowledge does not necessarily appear in data
- How to do that?
  - Add parts to the **loss function**, to have system respect constraints
  - Add **fake data points** to training, created with expert knowledge
  - Human-designed layers/modules (**differentiable programming**)

Multi-modal systems

- Able to accept inputs of different type: text, images, audio…
- Different pathways for each input
  - Feature extraction, like CNNs and others
  - **Feature fusion**, create features from multiple pathways
  - For a long time, it was a difficult problem

Multi-modal systems



Coding a vision language model from scratch!

Coding a Multimodal (Vision) Language Model from scratch in PyTorch with full explanation

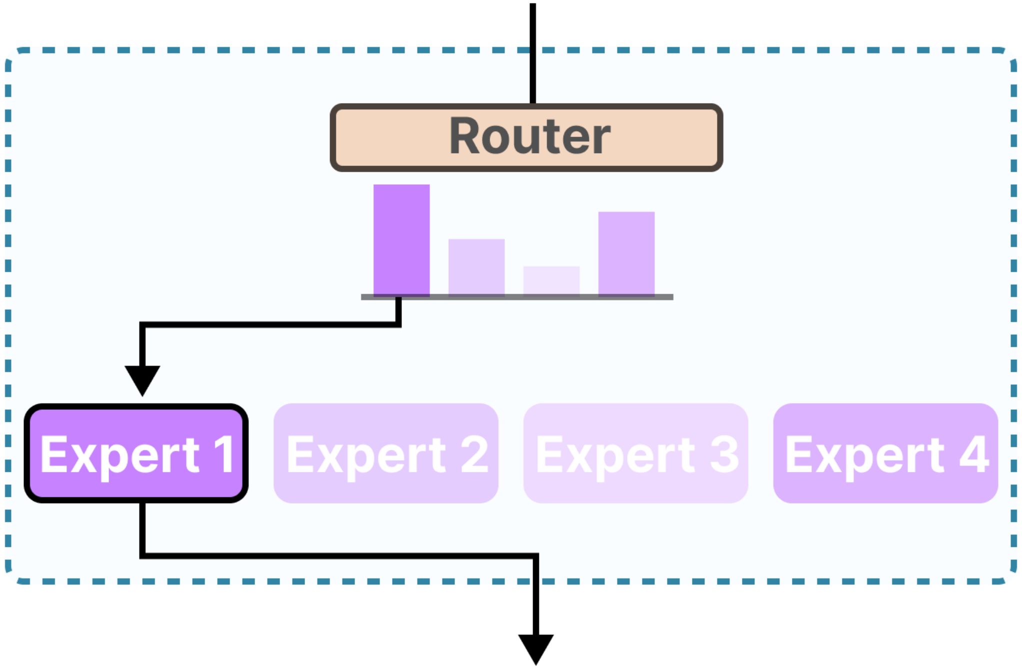
Umar Jamil
48.600 iscritti

50.598 visualizzazioni 4 mesi fa

https://youtu.be/vAmKB7iPkWw

Mixture of Experts (MoE)

- Using outputs of different specialized systems alternatively
- Softmax inside the network decides experts to activate

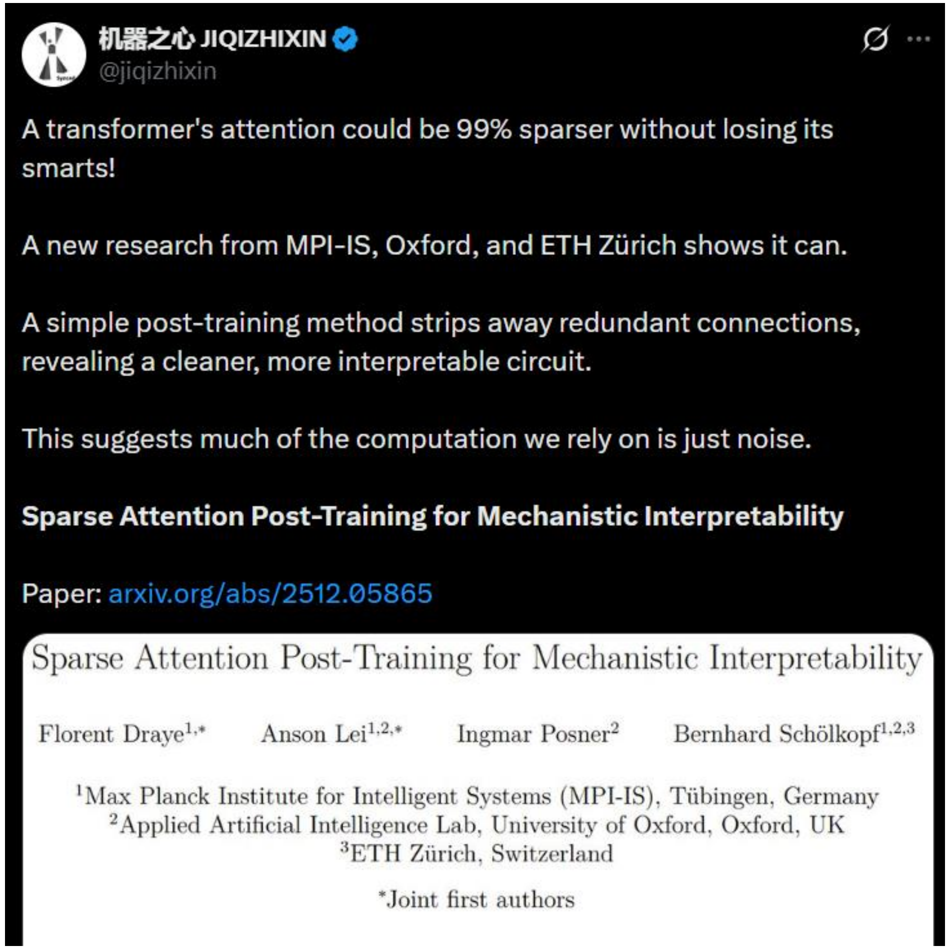


Large models on end-user machines

- Run large models (10^9 parameters) locally
  - Downsizing parameters (from float32 to float16…to int4 !!!)
  - Loss of performance, but works decently in inference (forward)
- Transfer learning on large models
  - Downsizing of parameters does not really work
  - Retrain a tiny amount of meaningful parameters
  - LoRA (Low-Rank Adaptation) seems to work well!
- Open-source versions of proprietary models (e.g. GPT-4)
- Sharing checkpoints and fine-tuned models

More efficiency

- In general MoE and downcasting improve **efficiency**
- Current LLMs are expensive for both training and inference
- Other solutions for efficiency: DeepSeek, sparsity



机器之心 JIQIZHIXIN
@jiqizhixin

A transformer's attention could be 99% sparser without losing its smarts!

A new research from MPI-IS, Oxford, and ETH Zürich shows it can.

A simple post-training method strips away redundant connections, revealing a cleaner, more interpretable circuit.

This suggests much of the computation we rely on is just noise.

Sparse Attention Post-Training for Mechanistic Interpretability

Paper: arxiv.org/abs/2512.05865

Sparse Attention Post-Training for Mechanistic Interpretability

Florent Draye[1,*] Anson Lei[1,2,*] Ingmar Posner[2] Bernhard Schölkopf[1,2,3]

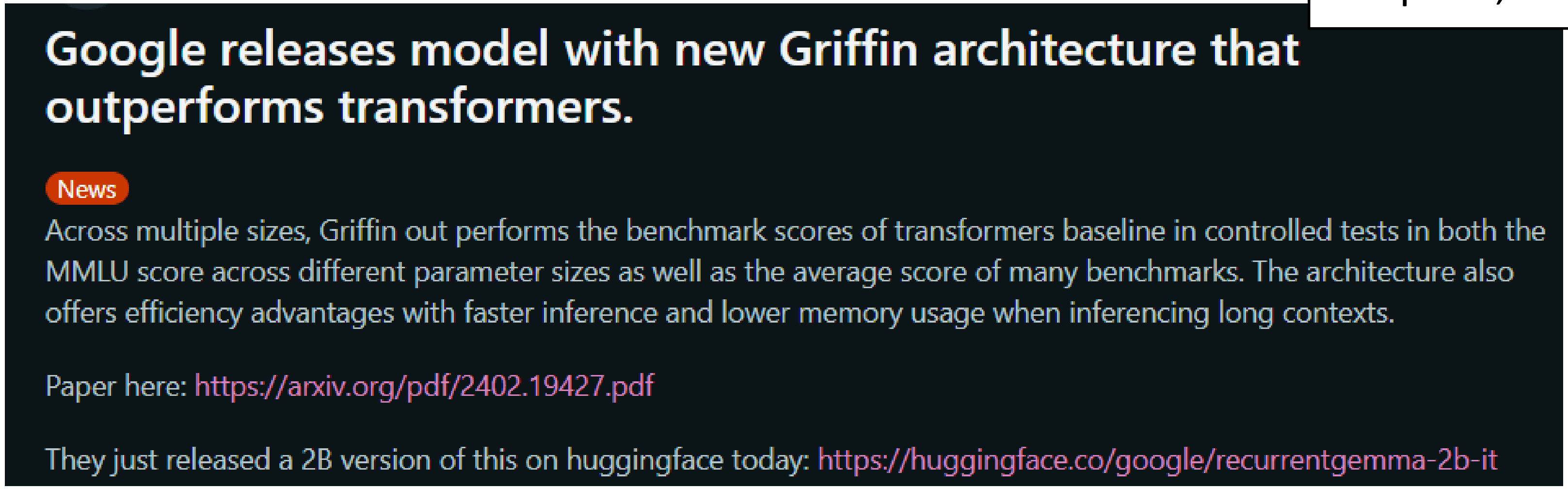
[1]Max Planck Institute for Intelligent Systems (MPI-IS), Tübingen, Germany
[2]Applied Artificial Intelligence Lab, University of Oxford, Oxford, UK
[3]ETH Zürich, Switzerland

[*]Joint first authors

New modules and architectures

- Transformers/Attention currently dominate the field
- But the race is far from over!
  - RMKV, large language model with RNN, https://www.rwkv.com/
  - Mamba, new types of modules for sequences
  - …

April 9, 2024



Google releases model with new Griffin architecture that outperforms transformers.

News

Across multiple sizes, Griffin out performs the benchmark scores of transformers baseline in controlled tests in both the MMLU score across different parameter sizes as well as the average score of many benchmarks. The architecture also offers efficiency advantages with faster inference and lower memory usage when inferencing long contexts.

Paper here: https://arxiv.org/pdf/2402.19427.pdf

They just released a 2B version of this on huggingface today: https://huggingface.co/google/recurrentgemma-2b-it

New models and architectures

- Diffusion/denoising/flow matching neural networks
  - Used for generation of images (but not only)
  - Idea: training data is generated by a **statistical distribution**
  - Learn how to **sample the distribution** to generate new data

- Using complex tricks from statistics
  - Approximate target distribution with Gaussian mixtures
  - Learn complex parameters with gradient descent

Retrieval Augmented Generation for LLMs

- Solution (?) to errors in LLM outputs
- Start from an established knowledge base
  - Search in the knowledge base, find documents from tags
  - Give documents in input to LLM
  - LLM generates summary based on documents

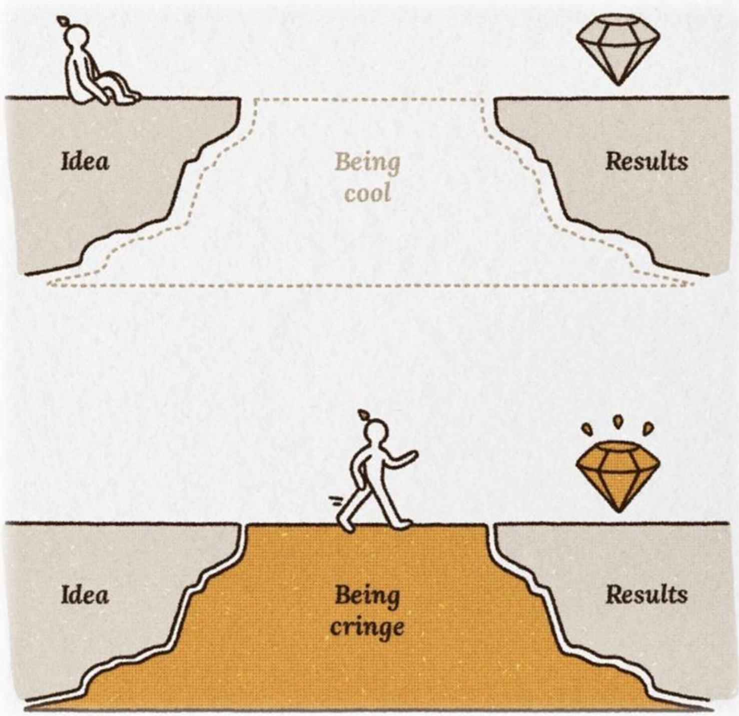
Chain of Thought / Reasoning

- It actually requires <u>no modifications</u> to loss or modules
  - Just modify the prompt of a LLM
  - Ask it to “reason step by step”
  - Surprisingly better effects
  - Already included in modern “reasoning” models

Are we all dead?

- AI is going to *replace every job* in the world!
- AI researchers will create a Machine God to rule over us!
- …

- The best take that I read
  - AI risks removing intermediate roles
  - “Good enough”
  - Without that step, no experts!



Are we all dead? (part 2)

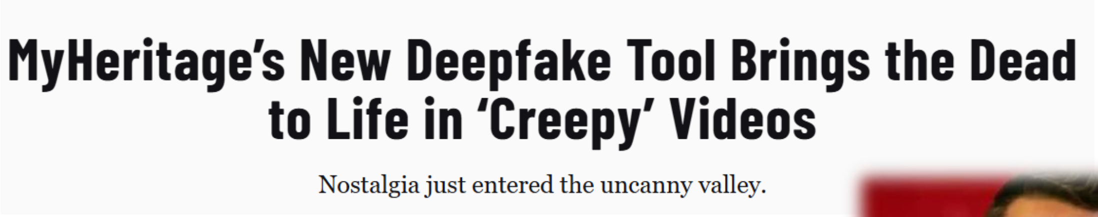
- Current AI is a bubble! It’s going to explode!
  - Big companies generate no profits
  - They are going to go bust when enthusiasm ends
  - This will create a massive economic shock

- This *could* be true
  - “Dot-com” bubble in 1995-2001
  - However, something good came out of it (eventually)

Necromancy

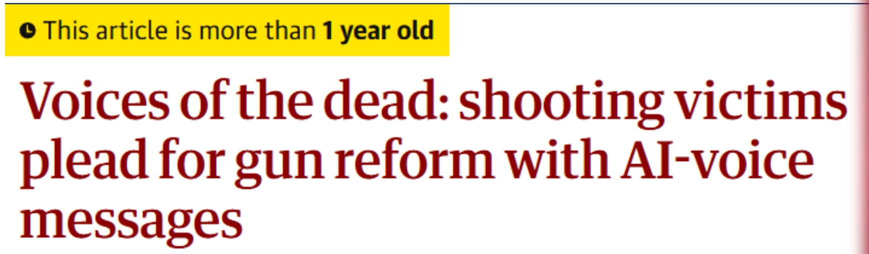
Necromancy

- Necromancy exists, and it is a subschool of computer science
  - Combination of image generation, cloning voice
  - LLM can emulate writing/speech style; basically a **ghost**
  - This is already happening



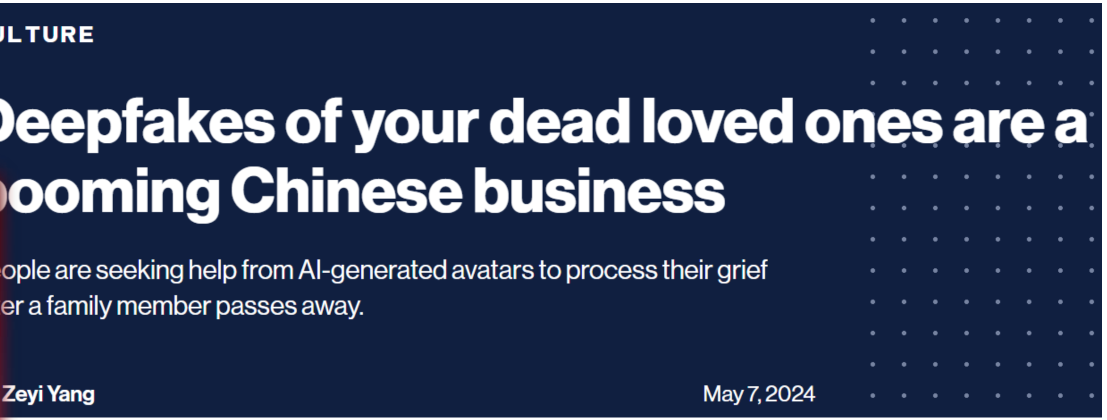
MyHeritage's New Deepfake Tool Brings the Dead to Life in 'Creepy' Videos

Nostalgia just entered the uncanny valley.



This article is more than **1 year old**

Voices of the dead: shooting victims plead for gun reform with AI-voice messages



CULTURE

Deepfakes of your dead loved ones are a booming Chinese business

eople are seeking help from AI-generated avatars to process their grief ter a family member passes away.

y Zeyi Yang

May 7, 2024

Beyond Deep Learning…?

- Two different world views
  1. All problems can be solved with **more/better data** and by **increasing the size** of the models
  2. In order to overcome current DL limitations, we need **radically different architectures** and taking into account **symbols**

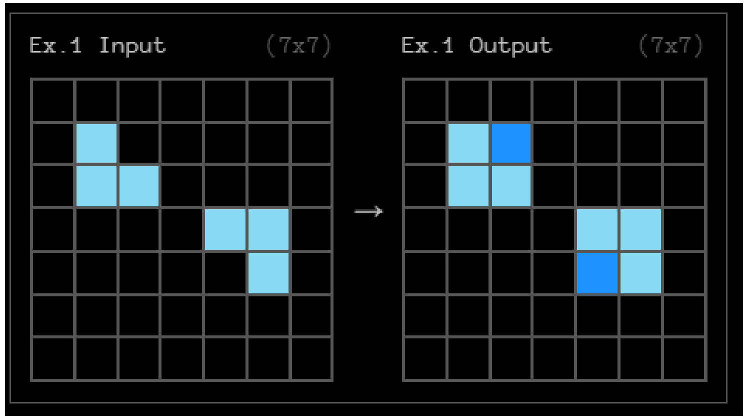
Questions?

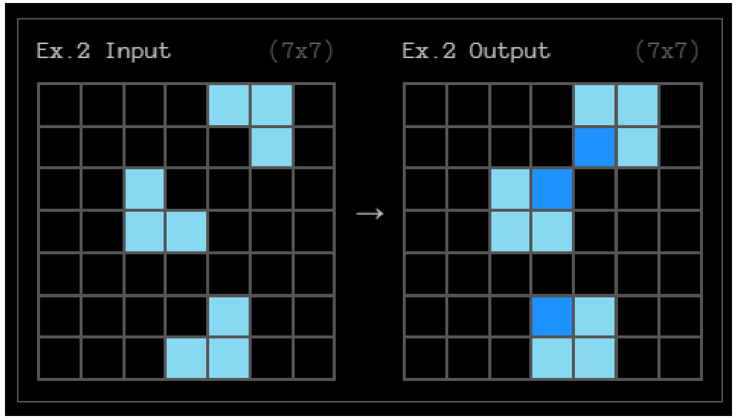
Bibliography

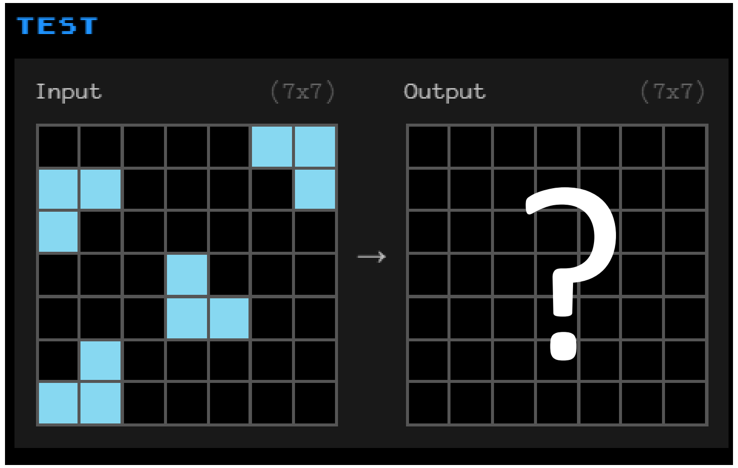
- ARC Prize Report, https://arcprize.org/2024-results
- Bergstrom C. & West J. *Calling Bullshit: Data Reasoning in a Digital World*, https://www.callingbullshit.org/
- Chollet, F. (2019). On the measure of intelligence. arXiv preprint arXiv:1911.01547.
- Cuomo, S. et al. (2022). *Scientific machine learning through physics–informed neural networks: Where we are and what's next*. Journal of Scientific Computing, 92(3), 88.
- De, S. et al. (2024). *Griffin: Mixing Gated Linear Recurrences with Local Attention for Efficient Language Models*. arXiv preprint arXiv:2402.19427.
- Marcus, G. (2020). *The next decade in AI: four steps towards robust artificial intelligence*. arXiv preprint arXiv:2002.06177.
- Gu, A., & Dao, T. (2023). *Mamba: Linear-time sequence modeling with selective state spaces*. arXiv preprint arXiv:2312.00752.
- Hu, E. J. et al. (2021). *LoRA: Low-rank adaptation of large language models*. arXiv preprint arXiv:2106.09685.
- Qu, X. et al. (2020). *Minimalistic attacks: How little it takes to fool deep reinforcement learning policies*. IEEE Transactions on Cognitive and Developmental Systems, 13(4), 806-817.
- Su, J., Vargas, D. V., & Sakurai, K. (2019). *One pixel attack for fooling deep neural networks*. IEEE Transactions on Evolutionary Computation, 23(5), 828-841.
- Wang, T. T. et al. (2023, July). *Adversarial policies beat superhuman go AIs*. In International Conference on Machine Learning (pp. 35655-35739). PMLR.
- Wu, X., & Zhang, X. (2016). *Automated inference on criminality using face images*. arXiv preprint arXiv:1611.04135, 4038-4052.
- Wu, X., & Zhang, X. (2016). *Responses to critiques on machine learning of criminality perceptions* (Addendum of arXiv: 1611.04135). arXiv preprint arXiv:1611.04135.

Images and videos: unless otherwise stated, I stole them from the Internet. I hope they are not copyrighted, or that their use falls under the Fair Use clause, and if not, I am sorry. Please don't sue me.

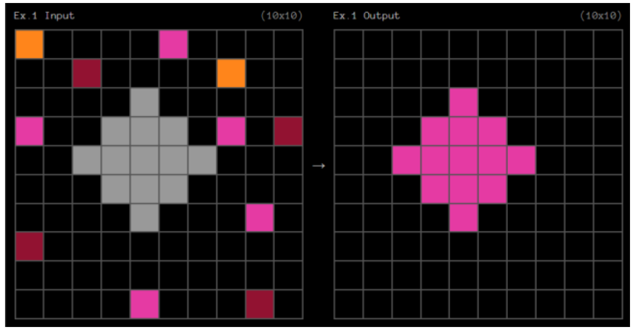
Beyond Deep Learning

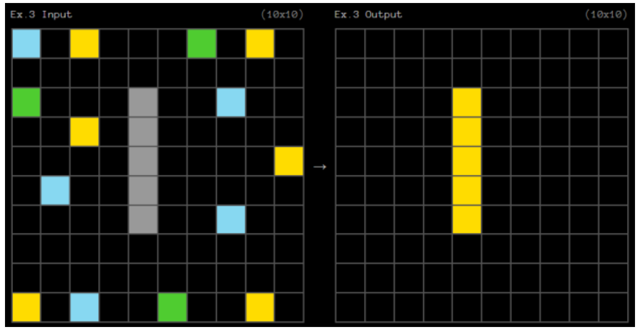


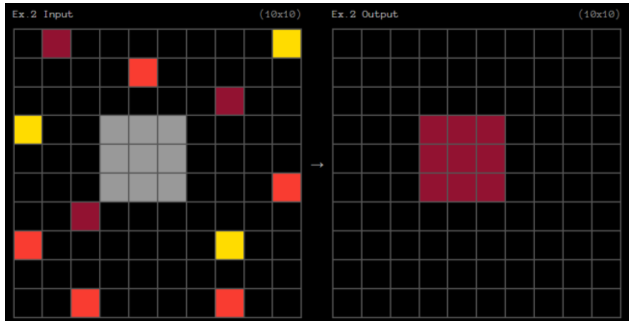


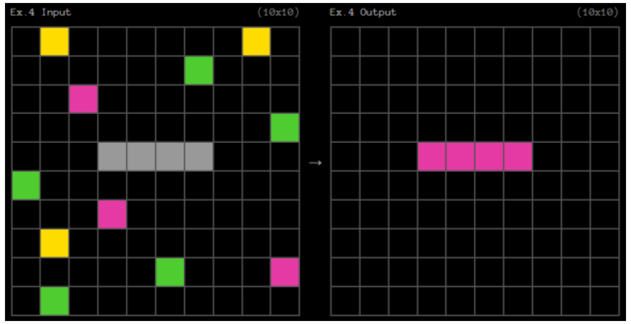


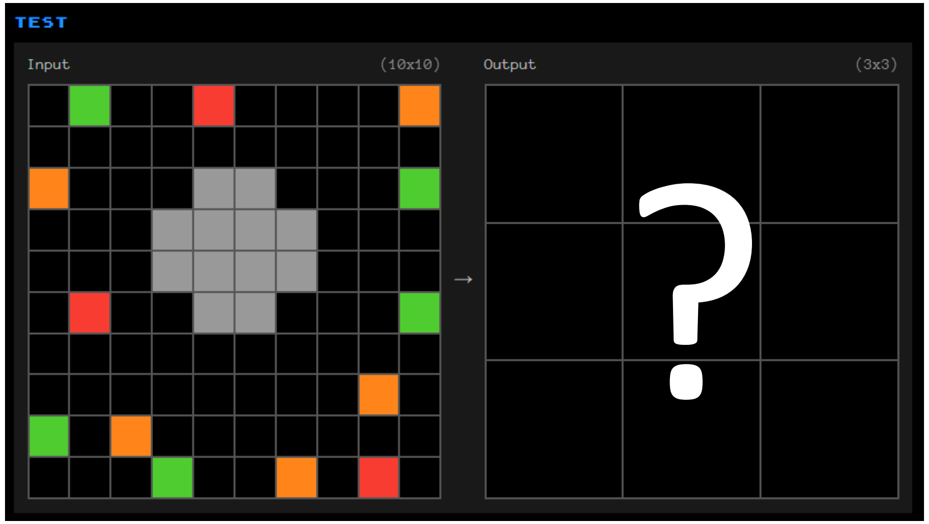
Beyond Deep Learning











Beyond Deep Learning…?

- DL/ML works well, but it is not **how humans learn**
  - Humans can learn from *few samples* and generalize well
  - Can research on human cognition help AI?
- "On the measure of intelligence", F. Chollet (2019)
  - Intelligence: ability to quickly and effectively learn new tasks
- Abstraction and Reasoning Corpus (ARC) Prize, 2024
  - Thousands of competitors, 1M$ prize (more or less)
  - Best result: 55% (vs human ~90%), https://arcprize.org/

Beyond Deep Learning…?

- Out of the competition, a *new idea*!
- Test-time tuning
  - A little bit like fine-tuning, but at **test time**
  - Slight modification of parameters based on test examples
  - Improves generalization to difficult cases
  - Requires some *data augmentation*