HUMANS AND AI

HUMAN INTELLIGENCE

- “Highly intelligent women have the tendency to marry less intelligent men.”
- Why is this the case?

HUMAN BRAIN VS. MACHINE LEARNING

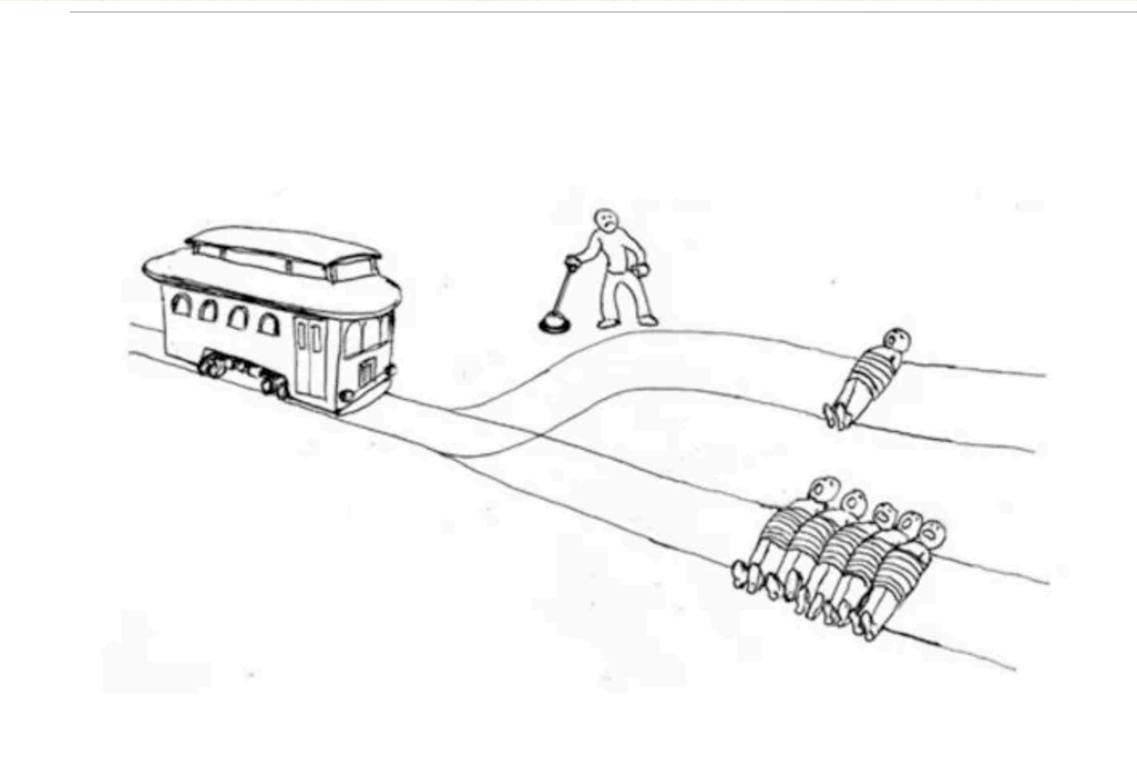
- Humans brains are …
  - causal (**plausible** explanations)
  - highly energy efficient
  - very slow
  - very bad at ignoring irrelevant features
  - influenced by emotions
  - Guided by inherent drives, reward function
- ML algorithms are …
  - statistical (**likely** explanations)
  - very energy hungry
  - very fast (and parallel)
  - great at feature selection
  - independent of emotions (which don’t exist)
  - Optimize clear objective function

CONCERNS WITH A.I.

WEAPONIZATION

- AI assisting humans with weapons
- AI “pulling the trigger” (Lethal Autonomous Weapons)
- AI can target and seek specific individuals

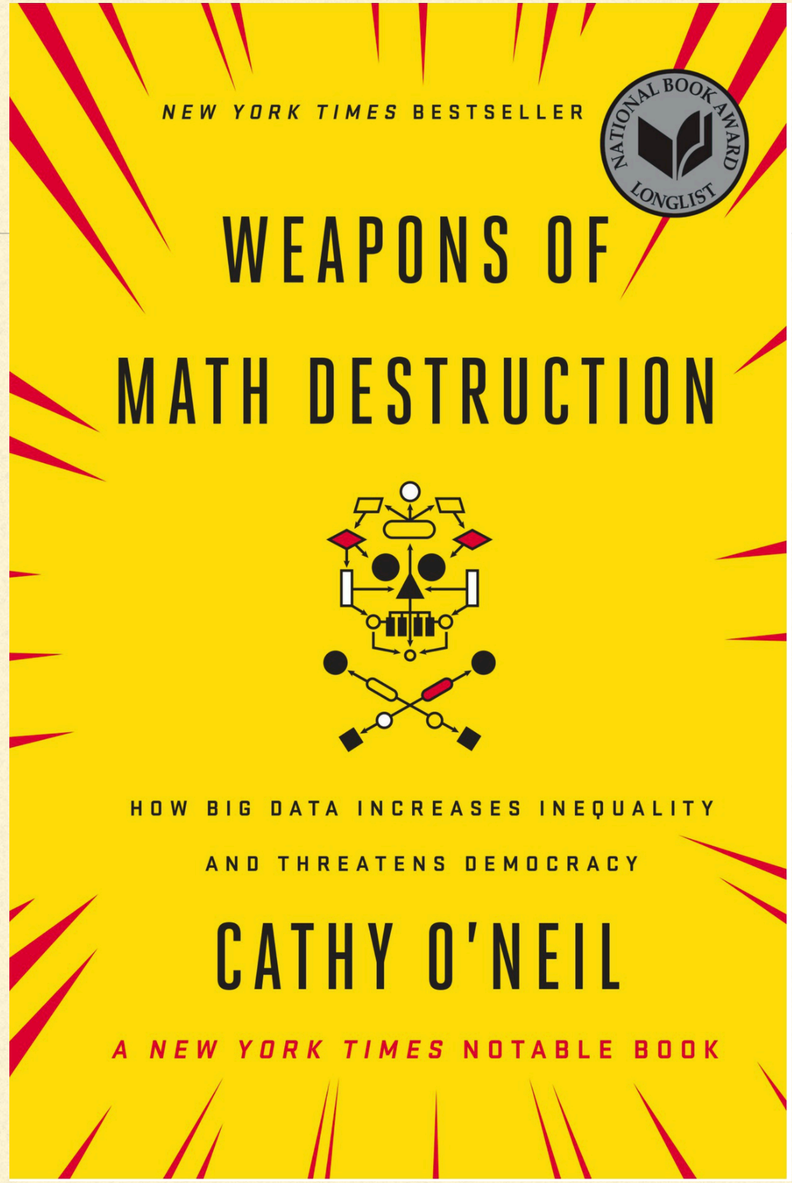
DILEMMAS

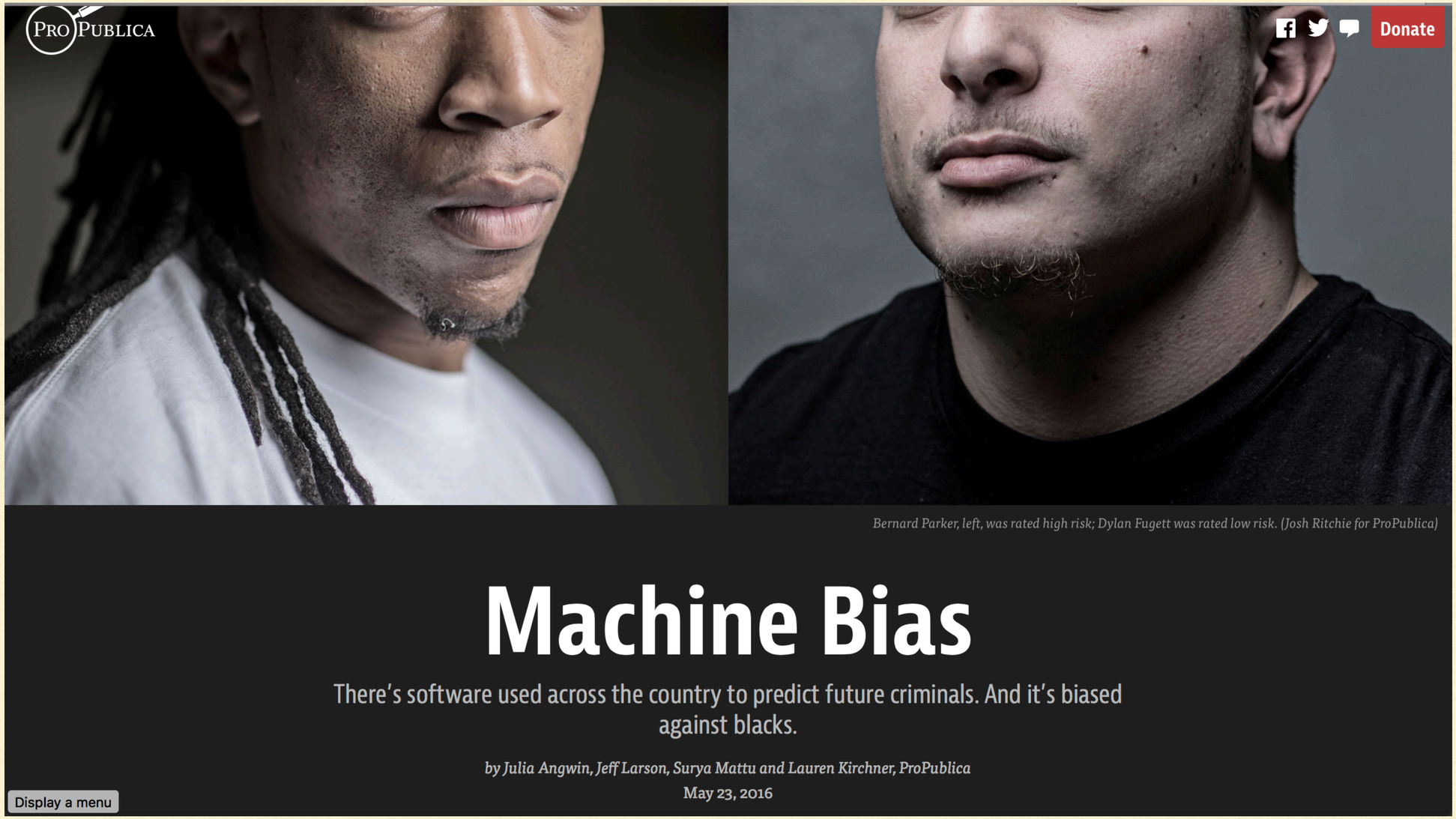


WMD

- Automated decision making (about humans)
  - Recidivism Prediction
  - Hiring
  - Performance evaluations
- Well intentioned:
  - Humans are slow and expensive
  - Humans are prejudiced / get tired / hangry

- Problems:
  - No **corrective** feedback loop
  - Biases in data are perpetuated
    - **Negative** feedback loops
  - Lack of transparency
  - Even rare errors are detrimental to victims
    - Multiple institutions using the same software, making the same errors / decisions





[https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing]

FAIRNESS

Inherent Trade-Offs in the Fair Determination of Risk Scores [J. Kleinberg, S. Mullainathan, M. Raghavan, ITCS 2016]

- Impossibility Theorem:

- **Fairness Properties for Risk Assignments.** Within the model, we now express the three conditions discussed at the outset, each reflecting a potentially different notion of what it means for the risk assignment to be "fair."

  (A) *Calibration within groups* requires that for each group t, and each bin b with associated score v_b, the expected number of people from group t in b who belong to the positive class should be a v_b fraction of the expected number of people from group t assigned to b.

  (B) *Balance for the negative class* requires that the average score assigned to people of group 1 who belong to the negative class should be the same as the average score assigned to people of group 2 who belong to the negative class. In other words, the assignment of scores shouldn't be systematically more inaccurate for negative instances in one group than the other.

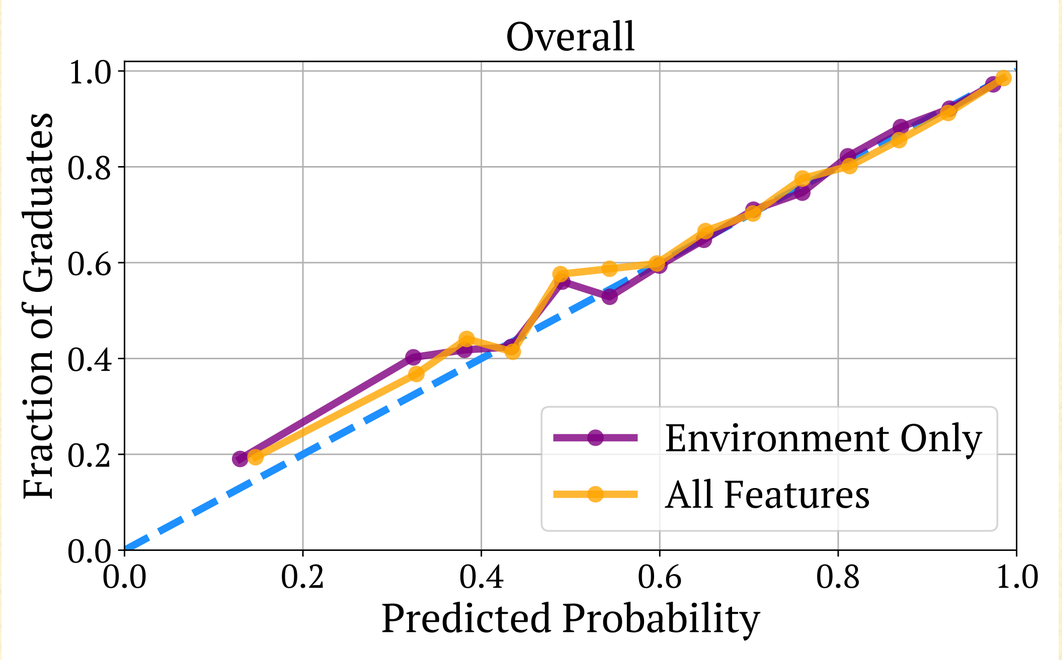
  (C) *Balance for the positive class* symmetrically requires that the average score assigned to people of group 1 who belong to the positive class should be the same as the average score assigned to people of group 2 who belong to the positive class.

- Not all three of these conditions can be satisfied unless the case is trivial.

WISCONSIN SCHOOL SYSTEM

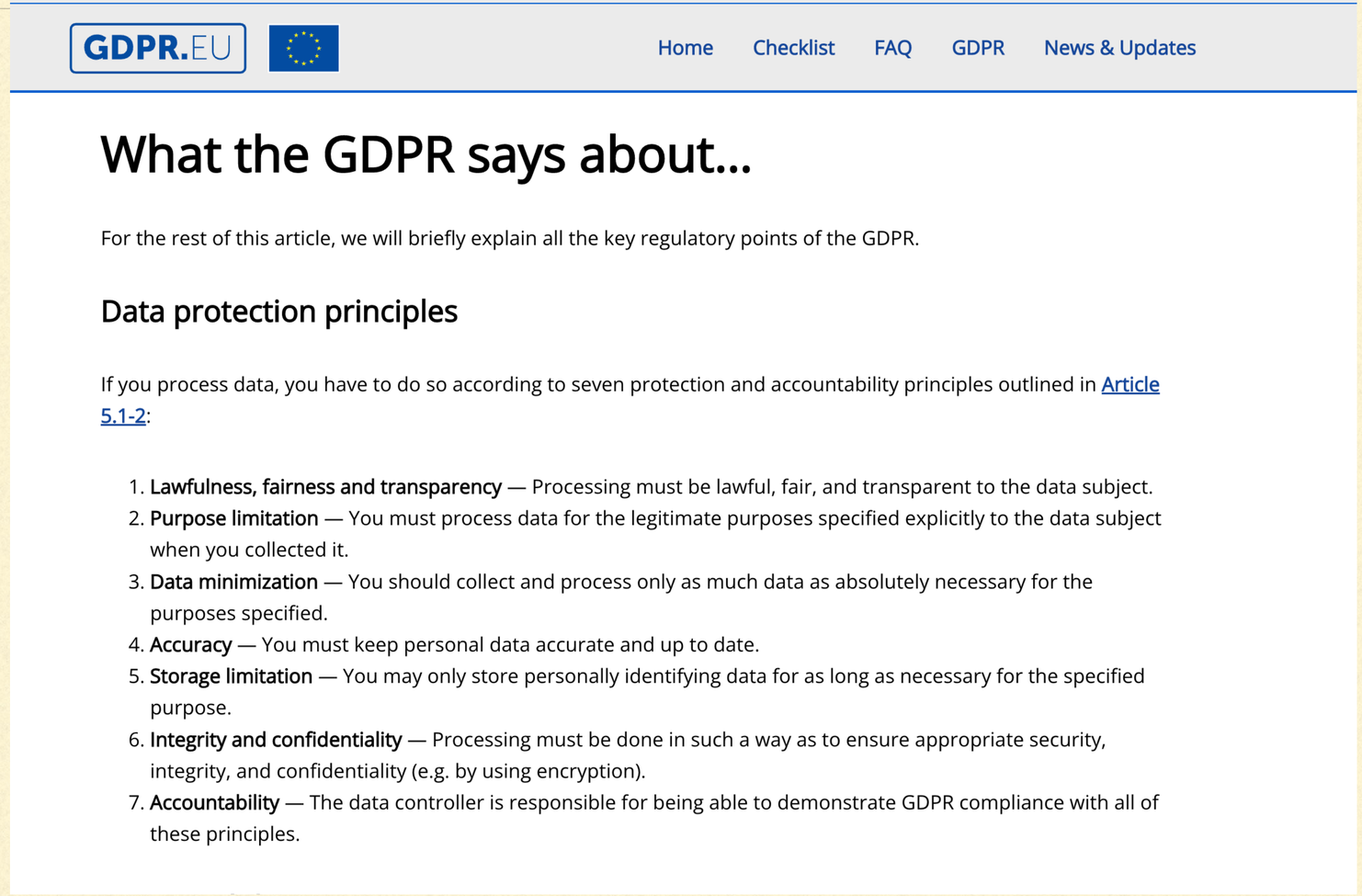
[Juan C. Perdomo, Tolani Britton, Moritz Hardt, Rediet Abebe]

- Early Intervention Prediction (i.e. will student drop out)
- Highly accurate on a state level
- Random on a within-school level



CONCERNS WITH DATA

- Companies collect data about individuals
- What about consent?
  - Withdraw consent? / “Right to be forgotten?”
- Protection against data leaks



GDPR.EU

Home Checklist FAQ GDPR News & Updates

What the GDPR says about...

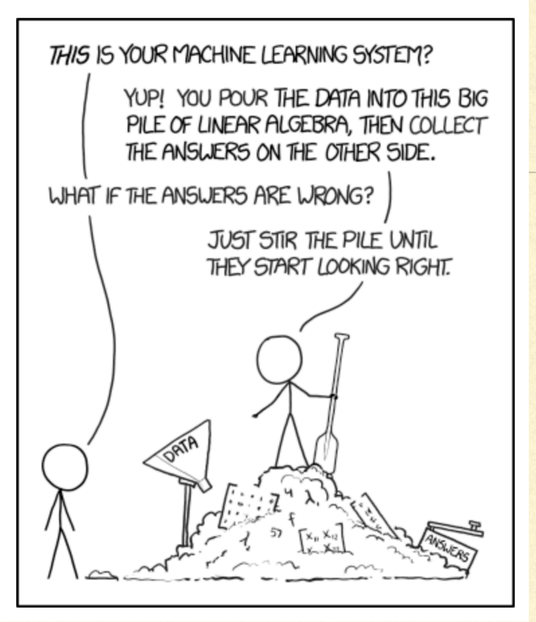
For the rest of this article, we will briefly explain all the key regulatory points of the GDPR.

Data protection principles

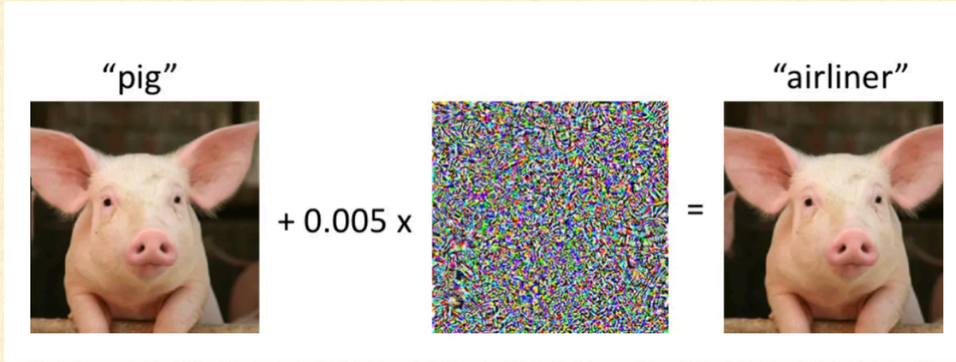
If you process data, you have to do so according to seven protection and accountability principles outlined in Article 5.1-2:

1. **Lawfulness, fairness and transparency** — Processing must be lawful, fair, and transparent to the data subject.
2. **Purpose limitation** — You must process data for the legitimate purposes specified explicitly to the data subject when you collected it.
3. **Data minimization** — You should collect and process only as much data as absolutely necessary for the purposes specified.
4. **Accuracy** — You must keep personal data accurate and up to date.
5. **Storage limitation** — You may only store personally identifying data for as long as necessary for the specified purpose.
6. **Integrity and confidentiality** — Processing must be done in such a way as to ensure appropriate security, integrity, and confidentiality (e.g. by using encryption).
7. **Accountability** — The data controller is responsible for being able to demonstrate GDPR compliance with all of these principles.

INTERPRETABILITY



- Small changes to input can make huge differences to output
- Very hard to obtain "certificates" of reasoning
- The more complex a model is, the harder it is to interpret
- Currently no satisfying solution for non-linear models

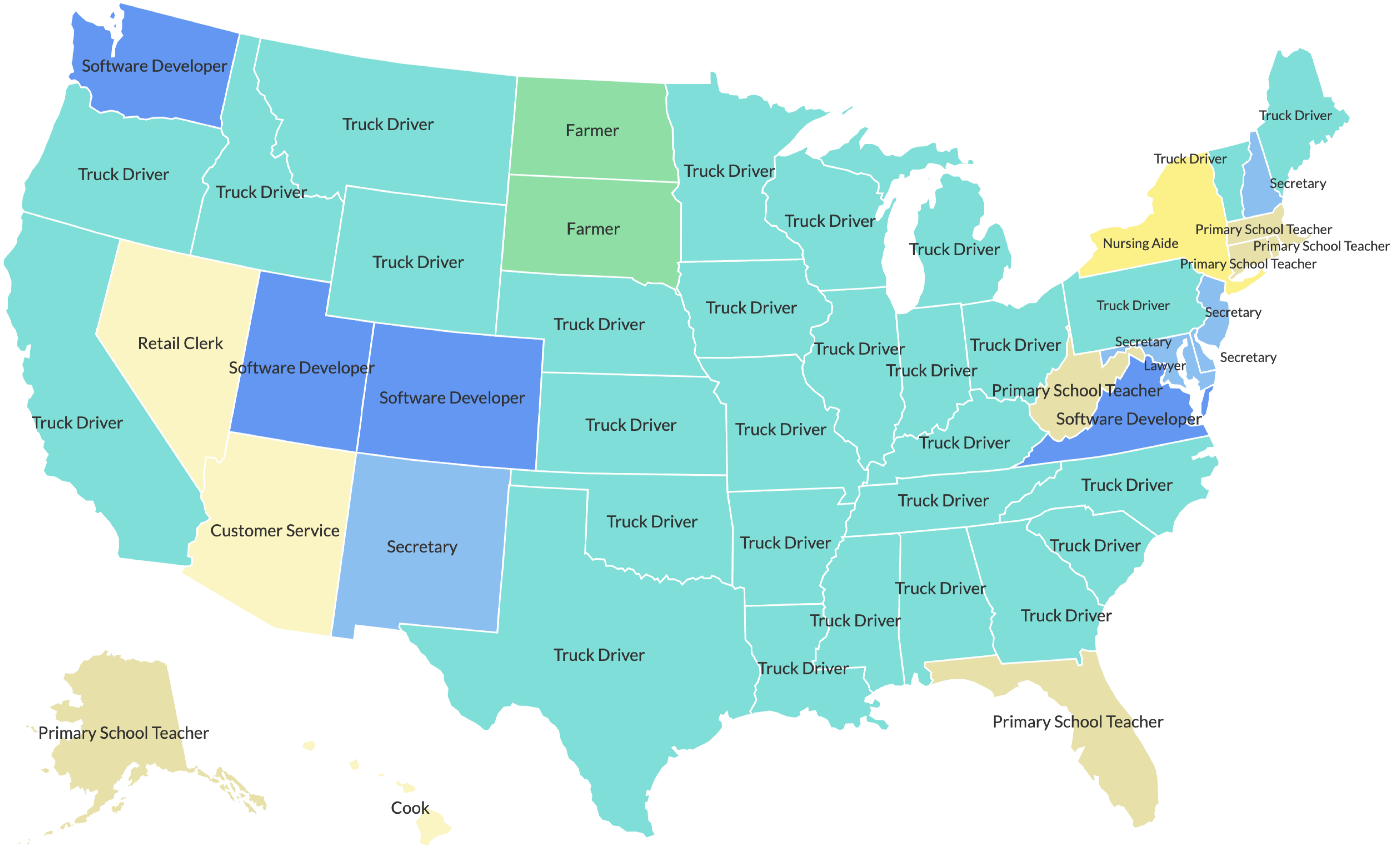


[https://medium.com/@smkirthishankar/the-unusual-effectiveness-of-adversarial-attacks-e1314d0fa4

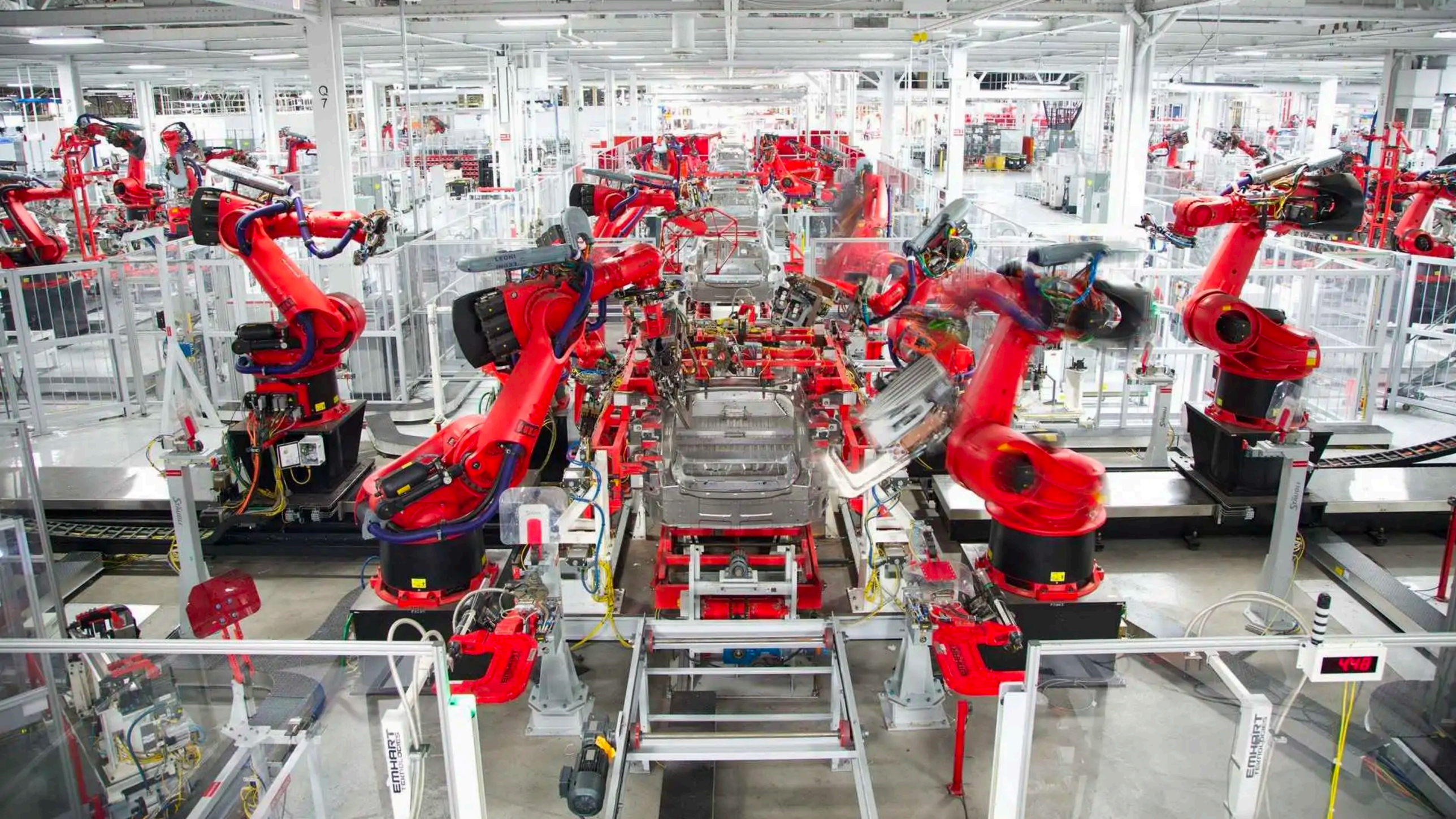
[Szegedy et al. 2013]

AI & JOBS

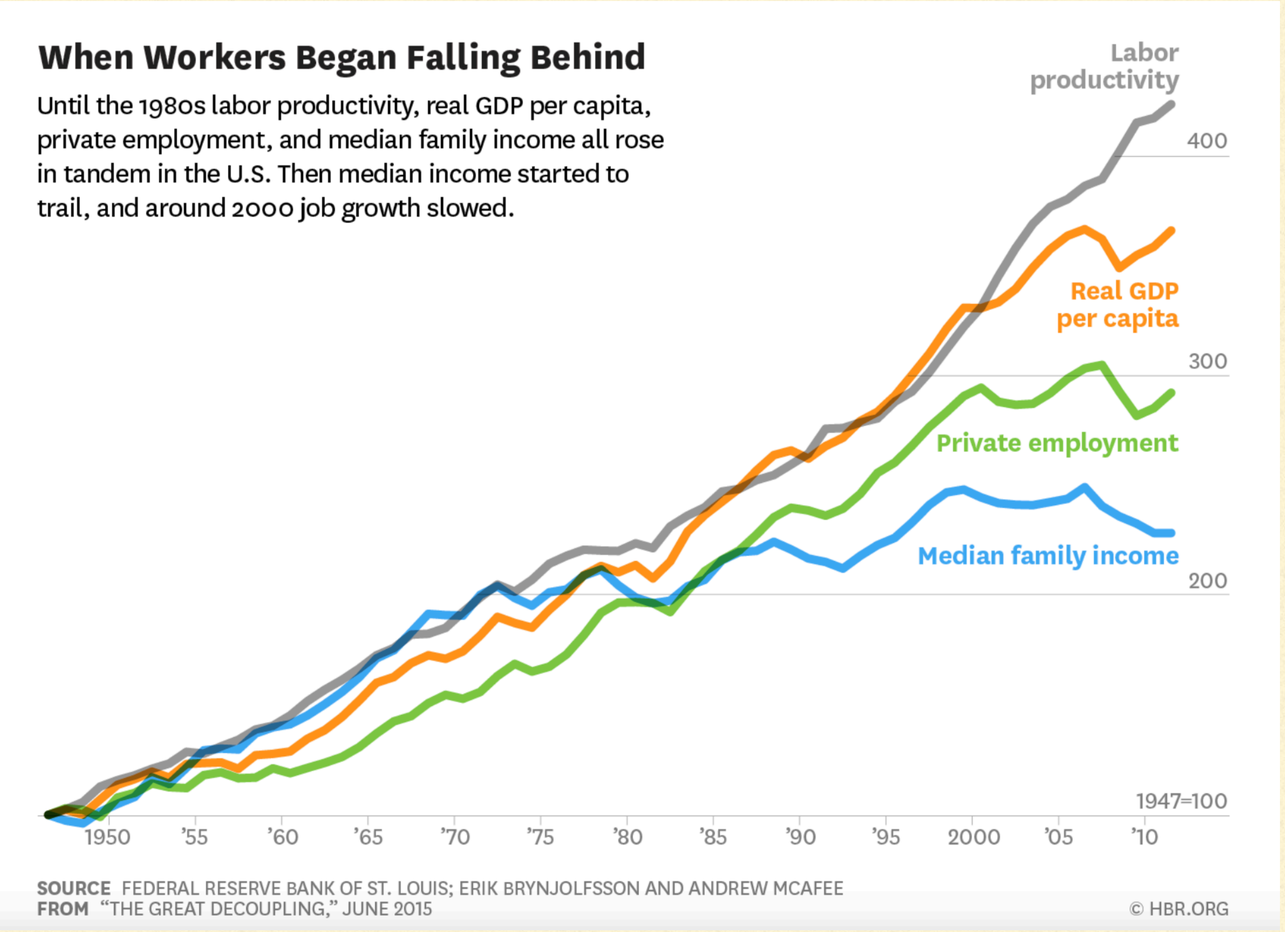
JOBS AND AI

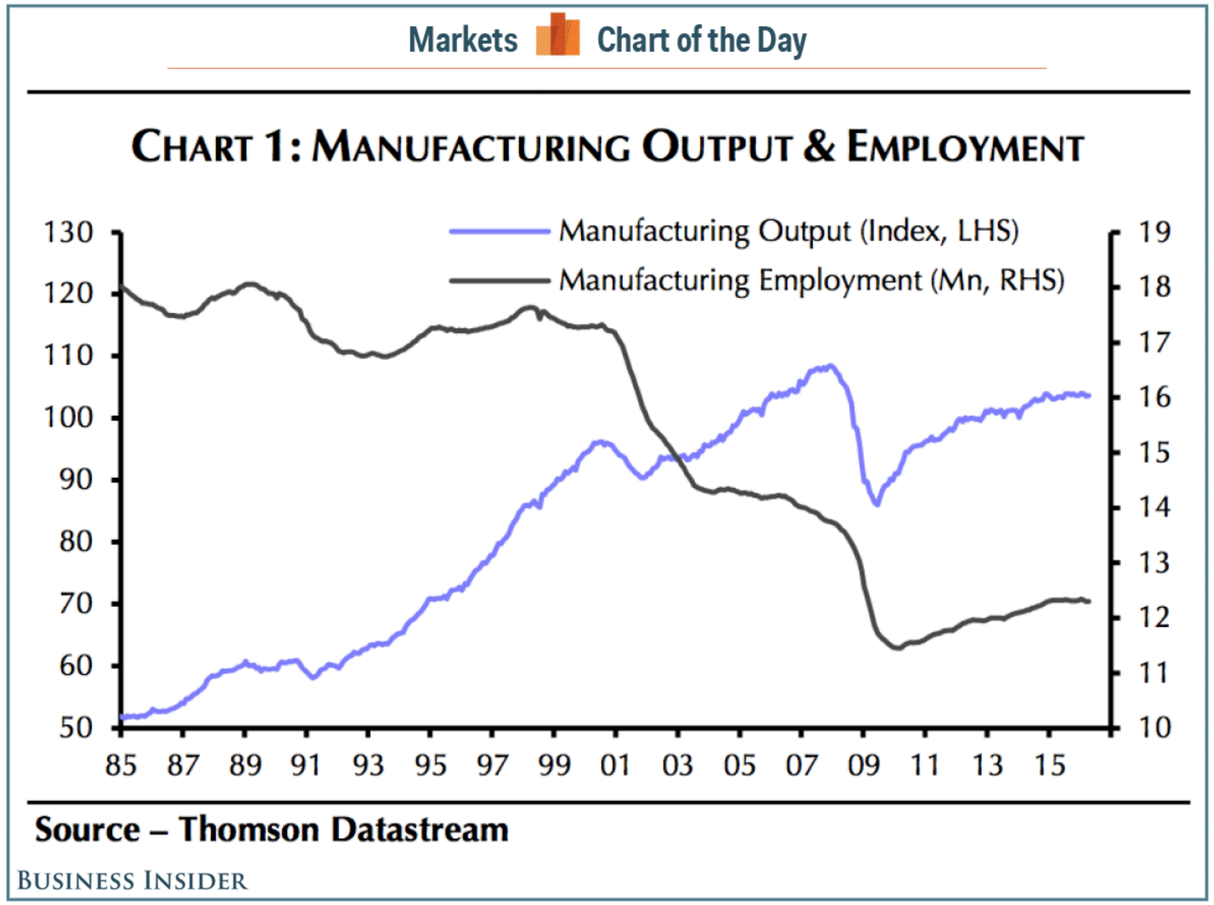


Most common jobs by state (2014)


EMHART
TECHNOLOGIES

EFFICIENCY IMPLIES JOBS?

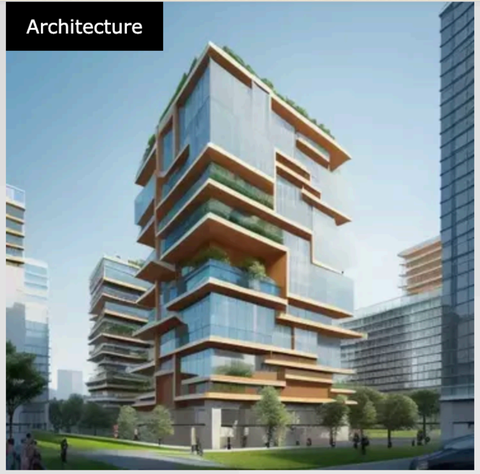


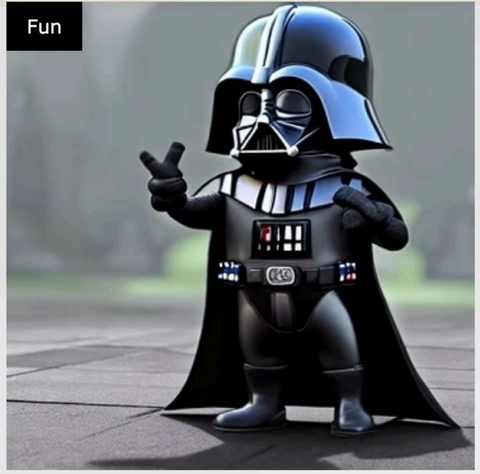


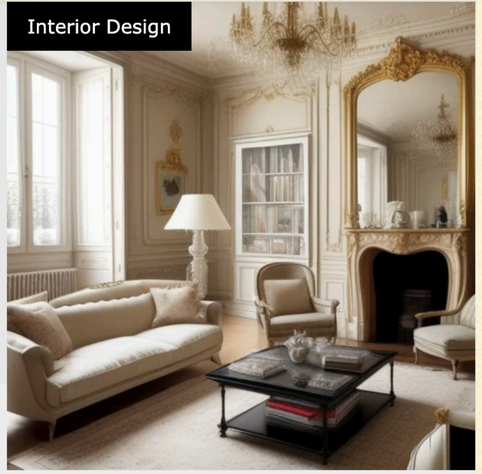
https://hbr.org/2015/06/the-great-decoupling

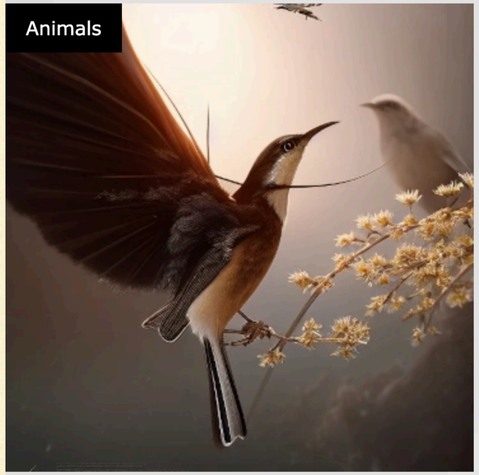
GENERATIVE A.I.


Art


Architecture


Fun


Interior Design

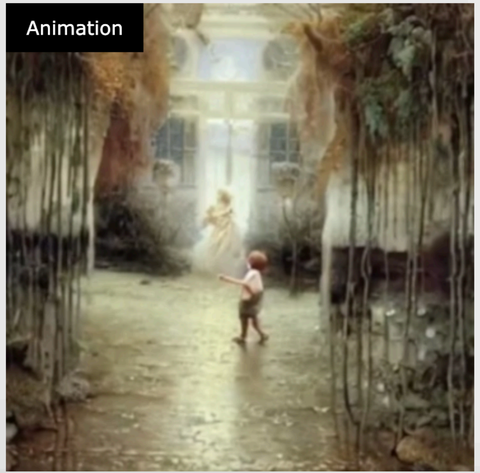

Animals

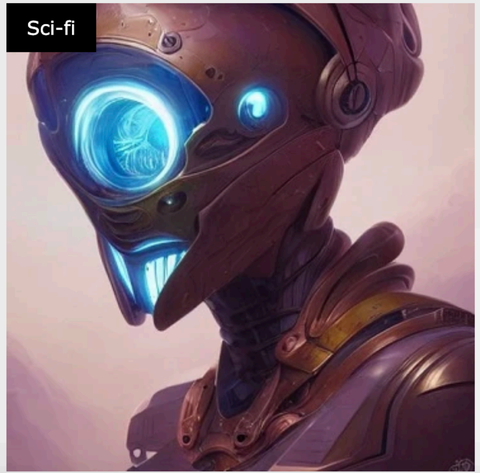

Seascape


Landscape


Fantasy


Misc


Animation


Sci-fi


Cats

“The Terror of War”, 1972 Nick Ut

Richard Nixon: ”I’m wondering if that was fixed"

https://upload.wikimedia.org/wikipedia/en/b/ba/The_Terror_of_War.jpg

Source: spiegel.de

Simulated exams	GPT-4 estimated percentile	GPT-4 (no vision) estimated percentile	GPT-3.5 estimated percentile
Uniform Bar Exam (MBE+MEE+MPT)[1]	298/400 ~90th	298/400 ~90th	213/400 ~10th
LSAT	163 ~88th	161 ~83rd	149 ~40th
SAT Evidence-Based Reading & Writing	710/800 ~93rd	710/800 ~93rd	670/800 ~87th
SAT Math	700/800 ~89th	690/800 ~89th	590/800 ~70th
Graduate Record Examination (GRE) Quantitative	163/170 ~80th	157/170 ~62nd	147/170 ~25th
Graduate Record Examination (GRE) Verbal	169/170 ~99th	165/170 ~96th	154/170 ~63rd
Graduate Record Examination (GRE) Writing	4/6 ~54th	4/6 ~54th	4/6 ~54th
USABO Semifinal Exam 2020	87/150 99th–100th	87/150 99th–100th	43/150 31st–33rd
USNCO Local Section Exam 2022	36/60	38/60	24/60
Medical Knowledge Self-Assessment Program	75%	75%	53%
Codeforces Rating	392 below 5th	392 below 5th	260 below 5th
AP Art History	5 86th–100th	5 86th–100th	5 86th–100th
AP Biology	5 85th–100th	5 85th–100th	4 62nd–85th
AP Calculus BC	4 43rd–59th	4 43rd–59th	1 0th–7th
AP Chemistry	4 71st–88th	4 71st–88th	2 22nd–46th
AP English Language and Composition	2 14th–44th	2 14th–44th	2 14th–44th
AP English Literature and Composition	2 8th–22nd	2 8th–22nd	2 8th–22nd
AP Environmental Science	5 91st–100th	5 91st–100th	5 91st–100th
AP Macroeconomics	5 84th–100th	5 84th–100th	2 33rd–48th
AP Microeconomics	5 82nd–100th	4 60th–82nd	4 60th–82nd
AP Physics 2	4 66th–84th	4 66th–84th	3 30th–66th

GENERATIVE AI AND THE LAW

- Section 230 : Internet Decency Act
  - “no provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider.”
- Does not apply to Generative AI!!

TECH · A.I.

‘The Godfather of A.I.’ just quit Google and says he regrets his life’s work because it can be hard to stop ‘bad actors from using it for bad things’

BY **PRARTHANA PRAKASH**

May 1, 2023 at 1:55 PM EDT

“These things are totally different from us,” he says. “Sometimes I think it’s as if aliens had landed and people haven’t realized because they speak very good English.”

Geoffrey Hinton was the pioneer behind some key concepts powering A.I. tools today.
COLE BURSTON—BLOOMBERG/GETTY IMAGES

Geoffrey Hinton is the tech pioneer behind some of the key developments in artificial intelligence powering tools like ChatGPT that millions of people are using today. But the 75-year-old trailblazer says he regrets the work he has devoted his life to because of how A.I. could be misused.

Most Popular

TECH

‘The Godfather of A.I.’ just quit Google and says he regrets his life’s work because it can be hard to stop ‘bad actors...

May 1, 2023

BY **PRARTHANA PRAKASH**

POLITICS

An upstate New York school may keep its Native American logo and ‘Warriors’ nickname—

Pause Giant AI Experiments: An Open Letter

We call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4.

Signatures
27565

Add your signature

PUBLISHED
March 22, 2023

AI systems with human-competitive intelligence can pose profound risks to society and humanity, as shown by extensive research[1] and acknowledged by top AI labs.[2] As stated in the widely-endorsed Asilomar AI Principles, *Advanced AI could represent a profound change in the history of life on Earth, and should be planned for and managed with commensurate care and resources*. Unfortunately, this level of planning and management is not happening, even though recent months have seen AI labs locked in an out-of-control race to develop and deploy ever more powerful digital minds that no one – not even their creators – can understand, predict, or reliably control.

12th April 2023
Last revision: 19th April 2023

Policymaking in the Pause

What can policymakers do *now* to combat risks from advanced AI systems?

Signatories

Signatories list slowed due to high demand

We have collected over 30,000 signatures and hope signatories will b to catch up. The high-profile signatures near the top of the list are all

Yoshua Bengio, Founder and Scientific Director at Mila, Turing Prize winn Montreal

Stuart Russell, Berkeley, Professor of Computer Science, director of the author of the standard textbook "Artificial Intelligence: a Modern Approac

Bart Selman, Cornell, Professor of Computer Science, past president of

Elon Musk, CEO of SpaceX, Tesla & Twitter

Steve Wozniak, Co-founder, Apple

WHERE WILL IT TAKE US?

It might kill us all!

- Evil actors will use A.I. for evil
- Allows few to control many
- LLM are already smarter than many humans
- Will lead to massive job losses
- A.I. will manipulate humans
- A.I. objectives likely not aligned with ours
- Smart A.I. can create even smarter A.I.

It will be great!!

- AI will amplify human abilities
- If we are smart enough to build it, we can control it
- Many new jobs will be created!
- GPT is nothing special
- A cat is way smarter than any LLM
- LLMs have no real understanding

WHERE IS MY SELF-DRIVING CAR?

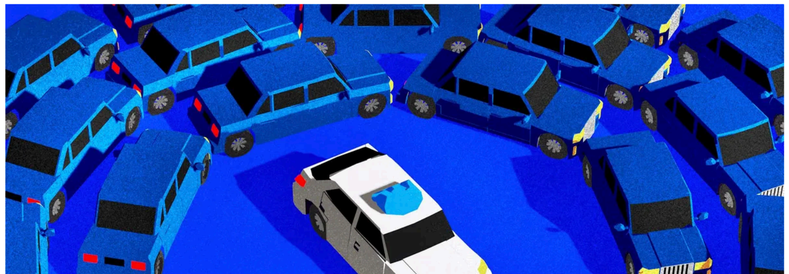
The Costly Pursuit of Self-Driving Cars Continues On. And On. And On.

Many in Silicon Valley promised that self-driving cars would be a common sight by 2021. Now the industry is resetting expectations and settling in for years of more work.

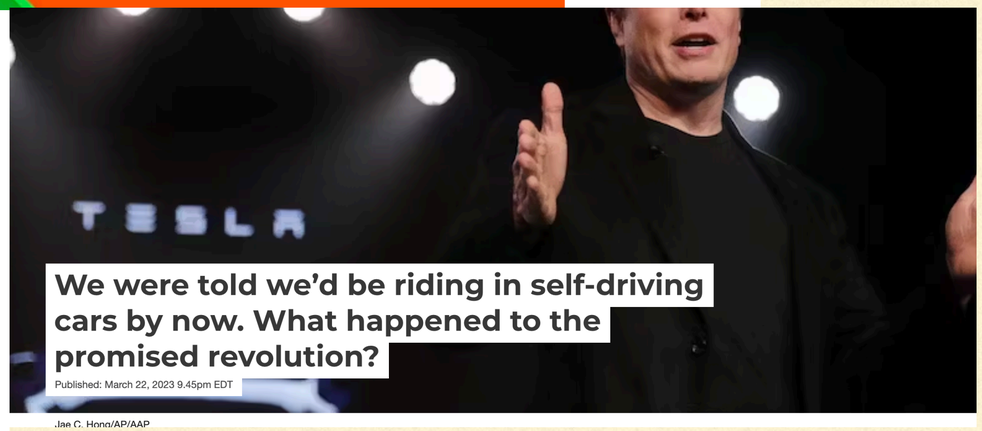
Give this article 329

TRANSPO / AUTONOMOUS CARS / CARS

Driverless cars aren’t going away, but we need to lower our expectations about them



/ The failure of Argo AI may lead some to speculate that autonomous vehicles are a lost cause. But the technology works, often very well; it just won’t be used in the way we originally



Jae C. Hong/AP/AAP

CONCLUSION

- AI promises wonderful, super-human capabilities but also potentially grave dangers
- The benefits could be unequally distributed
- Unintended (or intended) risks could spoil things for everybody
- Dangers are still poorly understood and could be underestimated

> Humanity can enjoy a flourishing future with AI. Having succeeded in creating powerful AI systems, we can now enjoy an "AI summer" in which we reap the rewards, engineer these systems for the clear benefit of all, and give society a chance to adapt. Society has hit pause on other technologies with potentially catastrophic effects on society.[5] We can do so here. Let's enjoy a long AI summer, not rush unprepared into a fall.

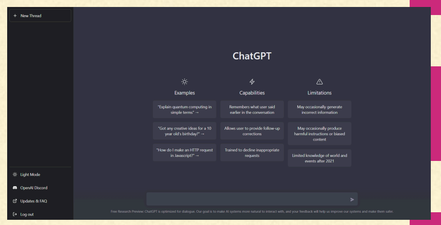
Intro to AI [CS4700]

Intro to ML [CS4780]

Fund. of ML [ECE4200]

Math Found. [CS4850]

Computer Vision [CS 4670]



Natural Language Processing [CS 4740]

Found. of Robotics [CS 4750]

Reinforcement Learning [CS 4789]

Math. Found. Of ML [CS 4783]

Principles of Large Sc. ML. [CS 4787]

Comp. Genetics [CS 4775]