



# The Age of Responsible AI has Begun

Prof. David Martens

Co-director ACRAI  
*Antwerp Center on Responsible AI*

*April 2024*

# Machine Learning

- **Machine Learning:** automatic extraction of knowledge from data
- Setting the scene with credit scoring example

Client	Income	Sex	Amount	Default
A	1.600	M	175.000	N
B	2.600	F	350.000	Y
C	3.280	M	50.000	N
D	950	M	120.000	Y
E	10.500	M	1.000.000	N
F	5.700	F	240.000	N
G	2.400	F	250.000	N

*Data*

Machine Learning

*Machine learning technique*

Classification Model
if income < 10.000 and Amount Loan > 100.000 and ... then default = yes

*Pattern*

Client	Income	Sex	Amount	Default
New client	2.000	F	500.000	Y

“99% of AI is prediction” Ng

# Terminology



**Mat Velloso**  
@matveloso



Difference between machine learning  
and AI:

If it is written in Python, it's probably  
machine learning

If it is written in PowerPoint, it's  
probably AI

# The Age of AI

## 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](#)

187k Accesses | 4797 Citations | 2936 Altmetric | [Metrics](#)

## The economic potential of generative AI: The next productivity frontier

June 14, 2023 | Report

## IMF BLOG

Artificial intelligence

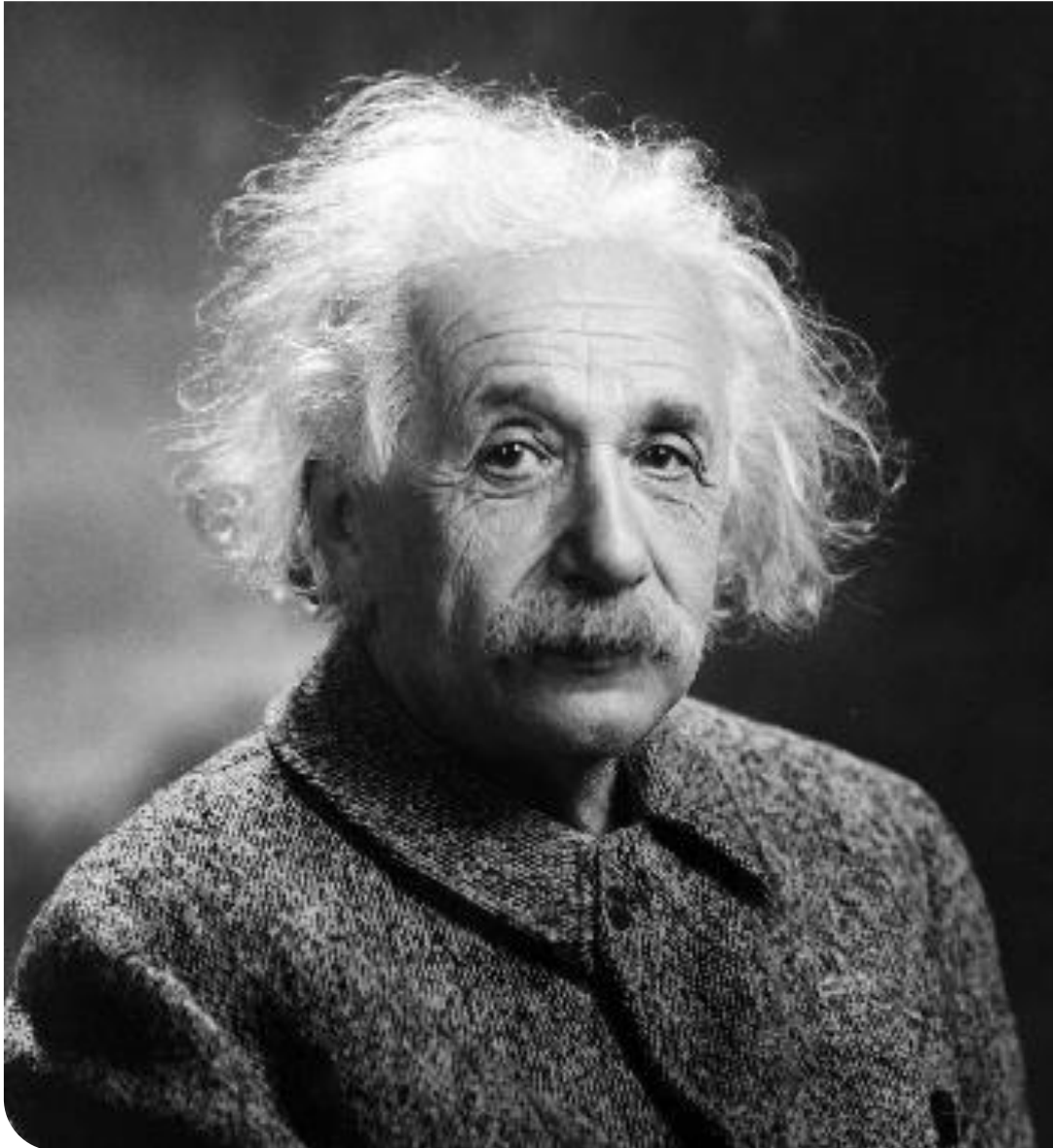
AI Will Transform the Global Economy. Let's  
Make Sure It Benefits Humanity. 

## GatesNotes

A NEW ERA

# The Age of AI has begun

By Bill Gates | March 21, 2023 • 14 minute read



## Einstein's Warning

*"Technological progress is like an axe in the hand of a pathological criminal."*

# Responsible AI

*The development and application of AI that is aligned with moral values of society*

# Why Care?

- Huge potential risks
- Data science ethics can bring value
- Expected from society
- AI Act is coming!

Data scientists and managers are not inherently unethical,  
but at the same time not trained to think this through neither

# AI Risks



Immediate



Systemic



Long term



# 1. Explainable AI



DHH ✓  
@dhh

Volgen

## Apple co-founder Steve Wozniak says Apple Card discriminated against his wife

By Clare Duffy, CNN Business  
Updated 16:15 GMT (00:15 HKT) November 11, 2019

The [@AppleCard](#) is such a fucking sexist program. It returns state time. I think it does

12:34 - 7

9.664 ret

1.585



why, but I swear we're not discriminating, IT'S JUST THE ALGORITHM". I shit you not. "IT'S JUST THE ALGORITHM!".

69 542 4.253



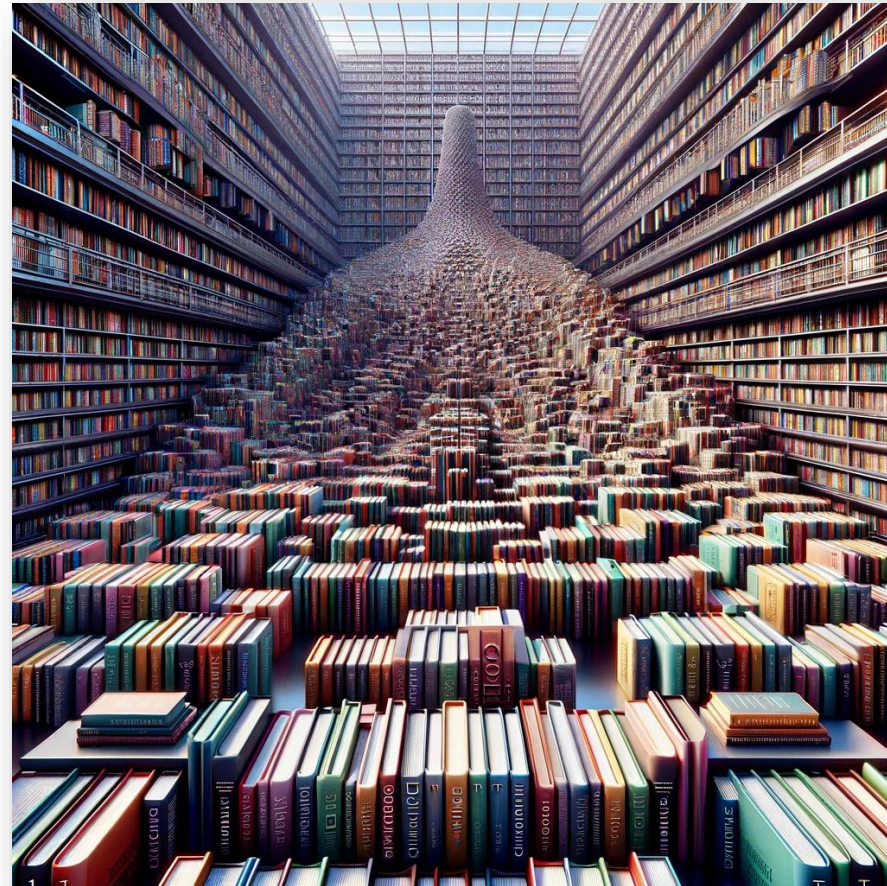
Financial Services is looking into the Apple Card, which is administered by

credit card accounts or any on Twitter, in reply to

direction though. It's big tech

# Black Box?

- Deep learning: large artificial neural network with massive number of parameters
  - MobileNetV2: 4.3 million parameters
  - GPT-4: >1 trillion parameters (an image of a printed version of the formula...)



# Trust: lab-setting versus real-life

- Data: image of skin lesion
- Task: diagnose skin cancer
- High test accuracy, matching accuracy of 21 dermatologists

[HTML] **Dermatologist-level classification of skin cancer with deep neural networks**

[A Esteva](#), [B Kuprel](#), [RA Novoa](#), [J Ko](#), [SM Swetter](#)... - nature, 2017 - nature.com

... of **deep learning** in dermatology, a technique that we apply to both general **skin** conditions and specific **cancers**. Using a single ... to **classify skin** lesions. The result is an algorithm that can ...

☆ Save 📄 Cite Cited by 10015 Related articles All 15 versions

# Instance-based explanations

**Example:** credit scoring using sociodemo and financial data



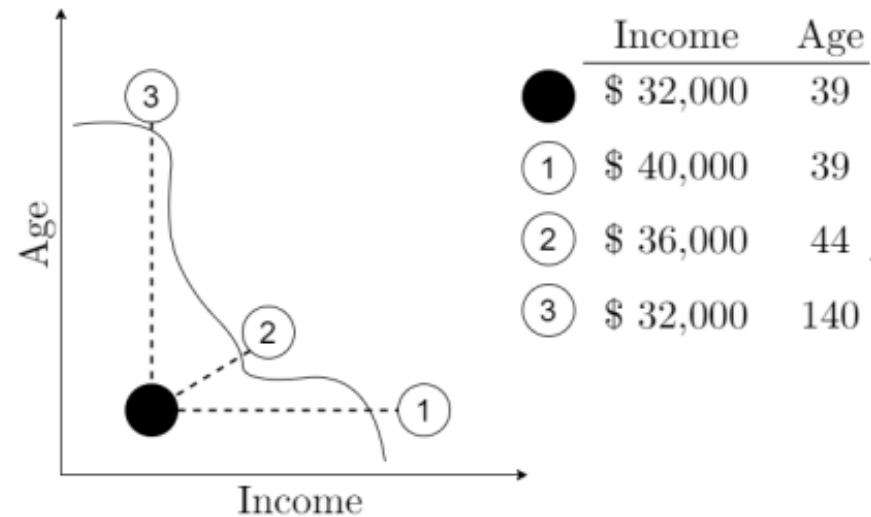
User  $x_i$ : Sam, with income \$ 32,000 and 39 years old.

Sam is denied credit

## WHY?

**IF** Sam would make 8,000\$ more

**THEN** his predicted class would change from *denied* to *granted*



# Imagine a world with AI explanations



Well, if your wife would also have had a 20+ year relationship with our bank, and would have been regarded as Premium customer at some point in time, she would also receive a 20x credit limit

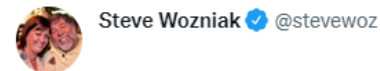


Well, if your wife's relationship status would have been "husband" instead of "wife", she would also receive a 20x credit limit

We clearly messed up, we're updating our models now.



Ah, ok, thanks for the additional feedback!



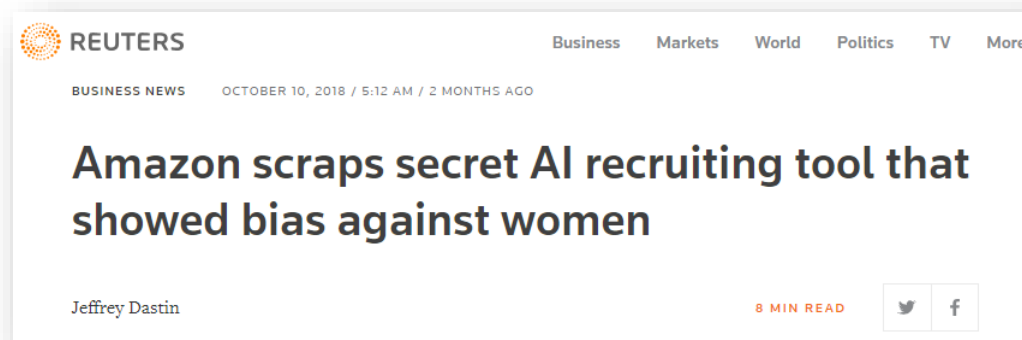
Glad you found this and react responsibly. It's how big tech should be in the 21st century.

## 2. Fairness

- Discrimination against sensitive groups
- Bias

## 2. Fairness

- Discrimination against sensitive groups
- Bias
  - HR Analytics, prediction model to review job applicants' resumes to automate the search for top talent
  - Trained on resumes from past (10 year period), biased data
  - Model trained to prefer male candidates, for example:
    - Penalized uses of word “woman’s” (eg woman’s chess club president)
    - Penalizes all-woman colleges



# What is fair?

- For example, in credit scoring:
  - An equal **proportion** of women and men are given credit
  - Of **those who apply**, an equal proportion of women and men are given credit
  - Of **those who apply and are qualified**, an equal proportion of women and men are given credit
  - The **accuracy** of the credit scoring model should be the same for women and men

Conflicting goals! Which one?



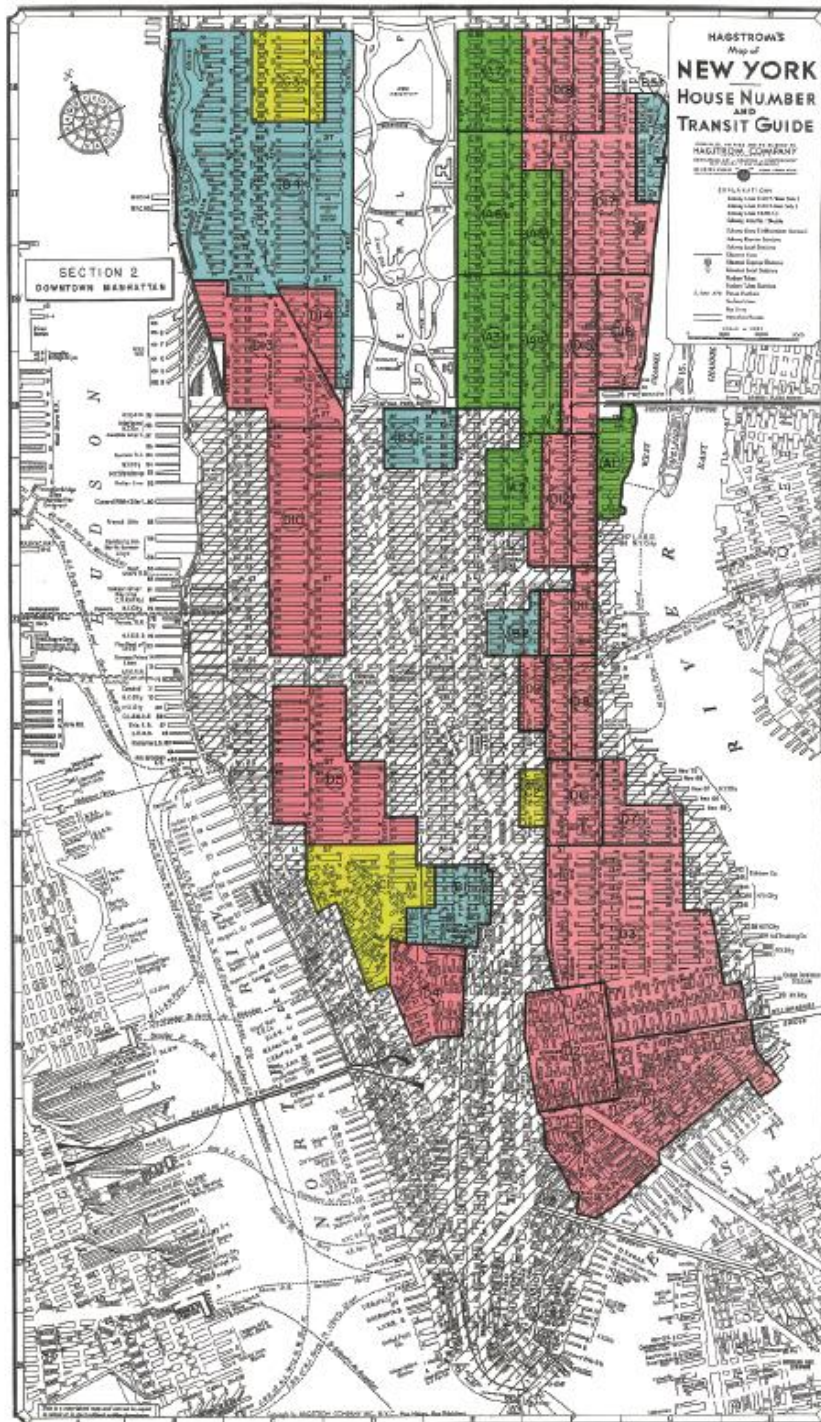


Fig. 4.7 HLOC security map of downtown Manhattan (New York) in 1932, where the red zones were considered 'hazardous' for lending companies. Taken from the *Mapping Inequality* project [378].

FORM 6  
10-1-32

AREA DESCRIPTION - SECURITY MAP OF New York City, N.Y.

1. AREA CHARACTERISTICS:

a. Description of Terrain-Level \_\_\_\_\_

b. Favorable Influences. None at present except city facilities - good transportation - adequate schools, etc. and easy access to Lower Manhattan business area.

c. Detrimental Influences. Age and obsolescence - present using 100% business or unimproved.

d. Percentage of land improved 100%; e. Trend of desirability next 10-15 yrs. Stable  
Est. 90%

2. INHABITANTS:

a. Occupation - labor - shop-workers; b. Estimated annual family income \$1000-2000

c. Foreign-born families 53%; Polish, Russian and Italian predominating; d. Negro No; %  
e. Infiltration of Italian; f. Relief families Many

3. Population is increasing; decreasing Yes; static

3. BUILDINGS:

	PREDOMINATING	or %	OTHER TYPE	10 %	OTHER TYPE	%
a. Type	1934-1941-family		40-50+ Years		Miscellaneous	
b. Construction	Brick				Brick	
c. Average Age	40+ Years					
d. Repair	Poor-to-poor				Fair	
e. Occupancy	60%		90%			
f. Home ownership	1%		Vertical		1%	
g. Constructed past yr.			None			
h. 1929 Price range	\$ 100\$		\$ 2000-3000		100\$	
i. 1932 Price range	\$ 1 \$ *		\$ *		\$ *	
j. 1935 Price range	\$ 1 \$ *		\$ *		\$ *	
k. Sales demand	\$		\$ Poor		\$	
l. Activity			Poor			
m. 1929 Rent range	\$ 100\$		\$ 2-10 per ft		100\$	
n. 1934 Rent range	\$ 1 \$ 2-10		" "		\$ *	
o. 1935 Rent range	\$ 1 \$ 2-10		" "		\$ *	
p. Rental demand	\$		\$ Good		\$	
q. Activity			Good			

4. AVAILABILITY OF MORTGAGE FUNDS: a. Home purchase None; b. Home building None

5. CLARIFYING REMARKS: Main business streets: 14th St., 1st Ave., Delancey St., Ave. A, Bowline, Rutgers, Houston, 2nd, 3rd Aves. See Area D-1. While this area is similar generally to the rest of the lower East Side it affords more opportunity for rehabilitation and replacement with low or moderate rent multi-dwellings. A section of the new East River Drive which is planned to extend ultimately the entire length of Manhattan, has been completed from 14th St. to the Williamsburg Bridge. This involved considerable demolition and the removal of a number of blocks. Units generally are in lowest brackets, although several modern apt. of quite good type constructed in the last 10 yrs. containing rentals of \$12-14 m.

6. NAME AND LOCATION: Lower East Side, New York City. SECURITY GRADE: D. AREA NO. 3

ASSESSED VALUES: Ratio to market price 125%

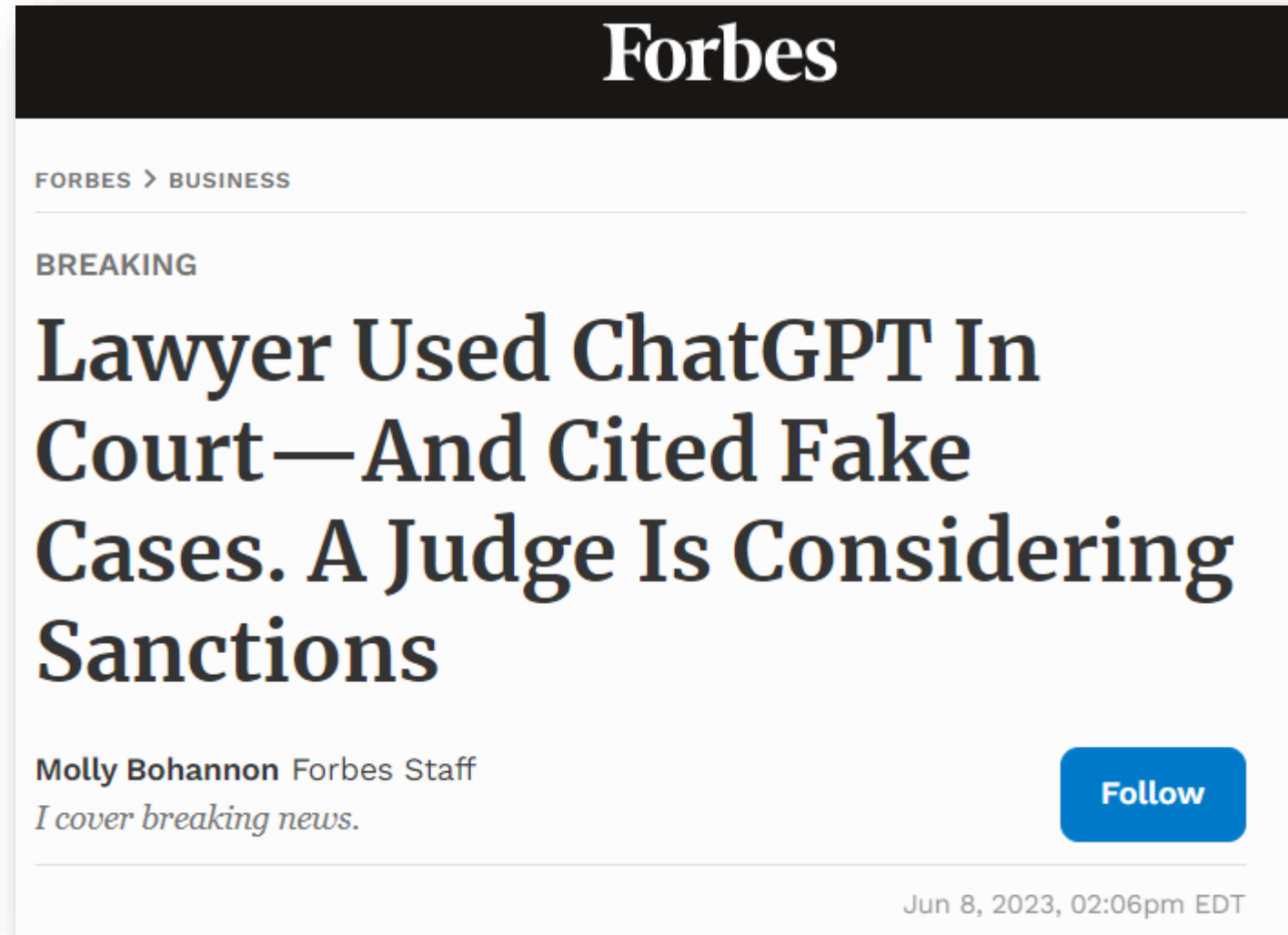
Area D3 - East Village: "c. Foreign-born families 53%; Polish, Russian and Italian predominating"  
Area D26 - Harlem: "d. Negroes: Yes, 90%"

# Redlining

- **Should banks never use location at all?**
  - If justified, based on economical motivations, such as:  
access to public transport, risk of flooding or earthquake.
- Might well correlate with race
- Be transparent in reasons and choice of fairness metric

## 3. Generative AI risks

- Hallucinations



The image is a screenshot of a Forbes article. At the top, the Forbes logo is displayed in white on a black background. Below the logo, the text 'FORBES > BUSINESS' is shown in a smaller, grey font. The word 'BREAKING' is written in a bold, black font. The main headline is 'Lawyer Used ChatGPT In Court—And Cited Fake Cases. A Judge Is Considering Sanctions', written in a large, bold, black font. Below the headline, the author's name 'Molly Bohannon Forbes Staff' and a bio 'I cover breaking news.' are displayed. A blue 'Follow' button is located to the right of the author's name. At the bottom right of the article, the date and time 'Jun 8, 2023, 02:06pm EDT' are shown.

**Forbes**

FORBES > BUSINESS

BREAKING

# Lawyer Used ChatGPT In Court—And Cited Fake Cases. A Judge Is Considering Sanctions

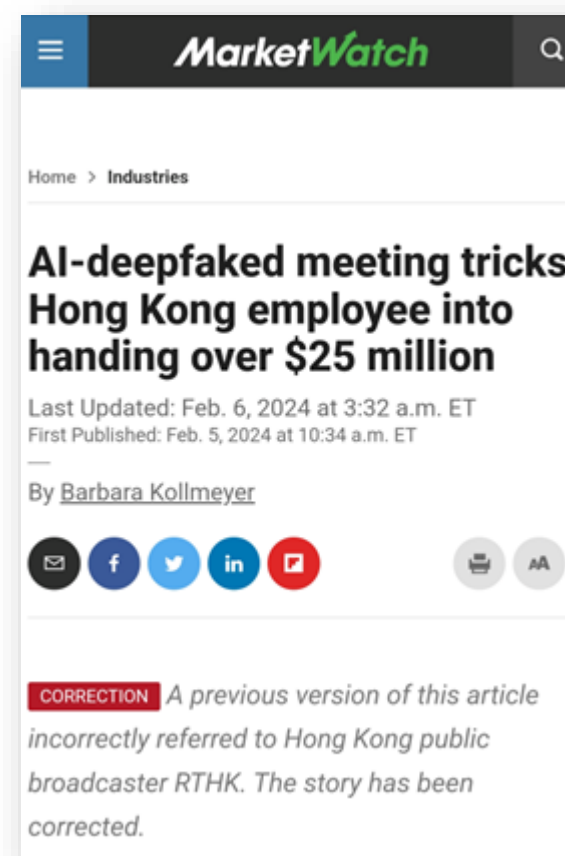
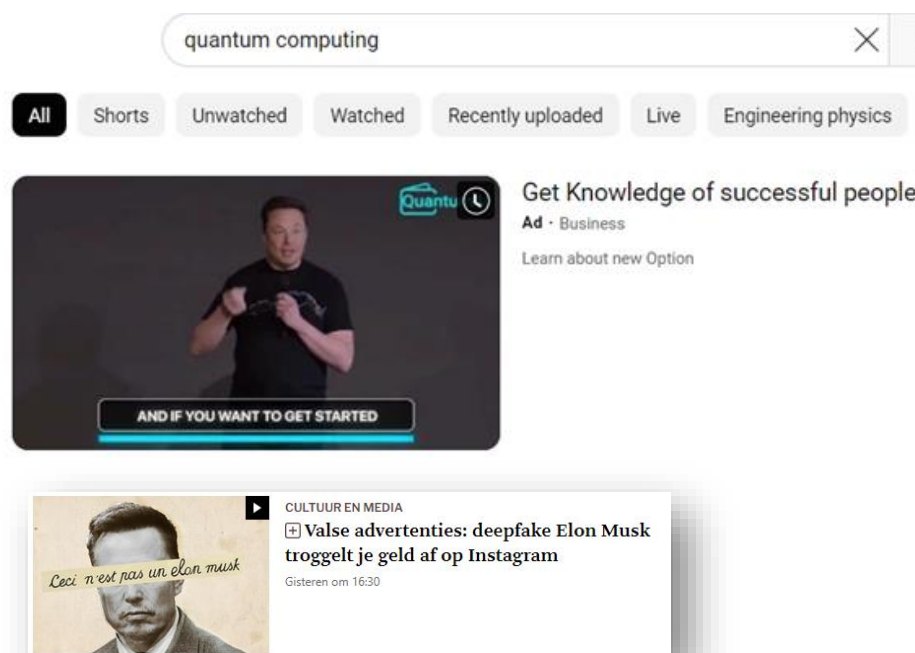
Molly Bohannon Forbes Staff  
*I cover breaking news.*

[Follow](#)

Jun 8, 2023, 02:06pm EDT

# Generative AI risks

1. Hallucinations
2. Misuse



# Should I use chatGPT?

- Study by Harvard, BCG on 758 consultants
- 18 realistic consulting tasks within the frontier of AI capabilities, consultants using AI:
  - 12.2% **more** tasks,
  - 25.1% more **quickly**,
  - significantly **higher quality** results
- For a task selected to be *outside* the frontier, 19% *less likely* to produce correct solutions compared to those without AI.

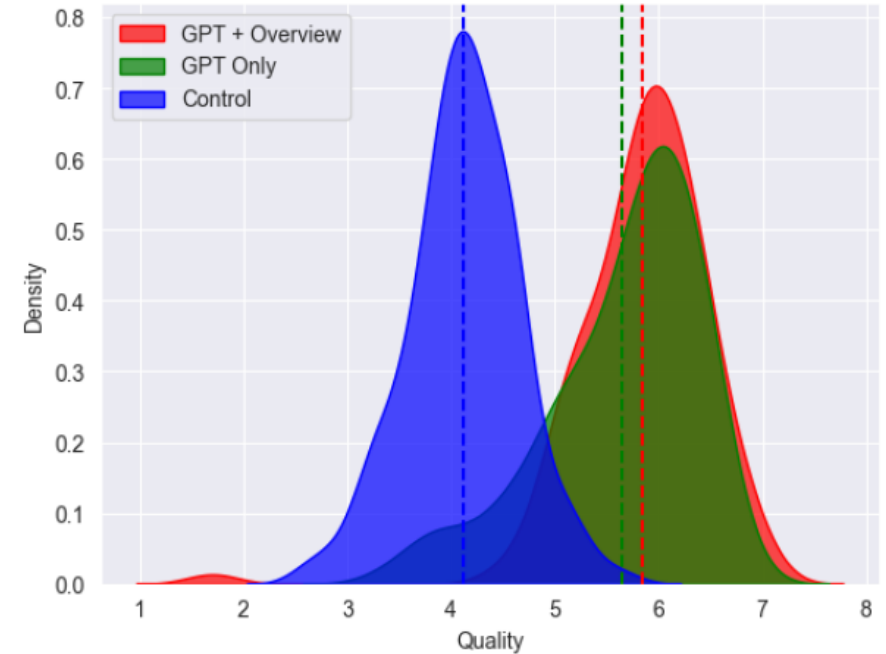
2023 WORKING PAPER HBS WORKING PAPER SERIES

## Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality

By: Fabrizio Dell'Acqua, [Edward McFowland III](#), Ethan Mollick, Hila Lifshitz-Assaf, Katherine C. Kellogg, Saran Rajendran, Lisa Kraymer, François Cadelon and [Karim R. Lakhani](#)

Format: Print | Language: English | Pages: 58

Figure 2: Performance Distribution - Inside the Frontier



Notes: This figure displays the full distribution of performance in the experimental task inside the frontier for subjects in the three experimental groups (red for subjects in the GPT+Overview condition; green for subjects in the GPT Only condition; blue for subjects in the control condition).

**Should I use chatGPT?**

**Absolutely, but do so responsibly!**

# AI Risks



Immediate



Systemic









Long term

# Systemic Risks

- **Concentration of power:** a few US and Asian giants dominating the field

Biggest AI companies in the world  
From sources across the web

 Microsoft	▼	 IBM	▼
 Google	▼	 OpenAI	▼
 Alphabet	▼	 SenseTime	▼

About 10% academics  
About 10% European





# Systemic Risks

- **Concentration of power:** a few US and Asian giants dominating the field
- Why is that bad?
  - They understand the technology (cf. discussion on open source)
  - They control the technology (cf. Elon Musk and Starlink)
  - They drive the discussions and regulations on what is right and wrong (cf. closed door meeting in Washington)

## Viewpoint: Europe needs a CERN for artificial intelligence

24 Oct 2023 | Viewpoint

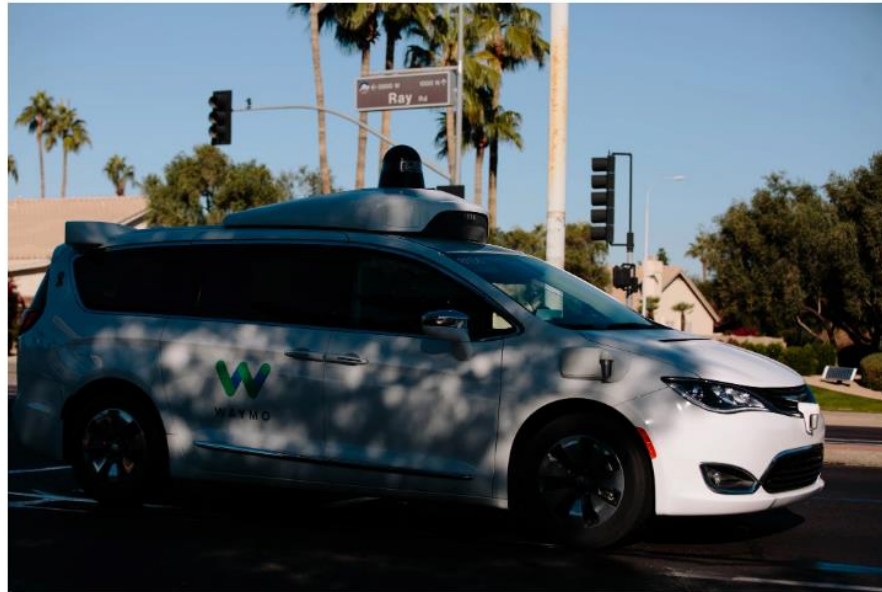
*Since the 'AI made in Europe' strategy launched in February 2020, the US has pulled further ahead. The EU's problem is a lack of scale and focus. The answer is to adopt CERN's approach to running large, coordinated and highly ambitious projects*

By [Holger Hoos](#) and [Morten Irgens](#)

# Systemic Risks

- **Jobs:** displacement and new opportunities

## *Wielding Rocks and Knives, Arizonans Attack Self-Driving Cars*



A Waymo autonomous vehicle in Chandler, Ariz., where the driverless cars have been attacked by residents on several occasions. Caitlin O'Hara for The New York Times

# AI Risks



Immediate



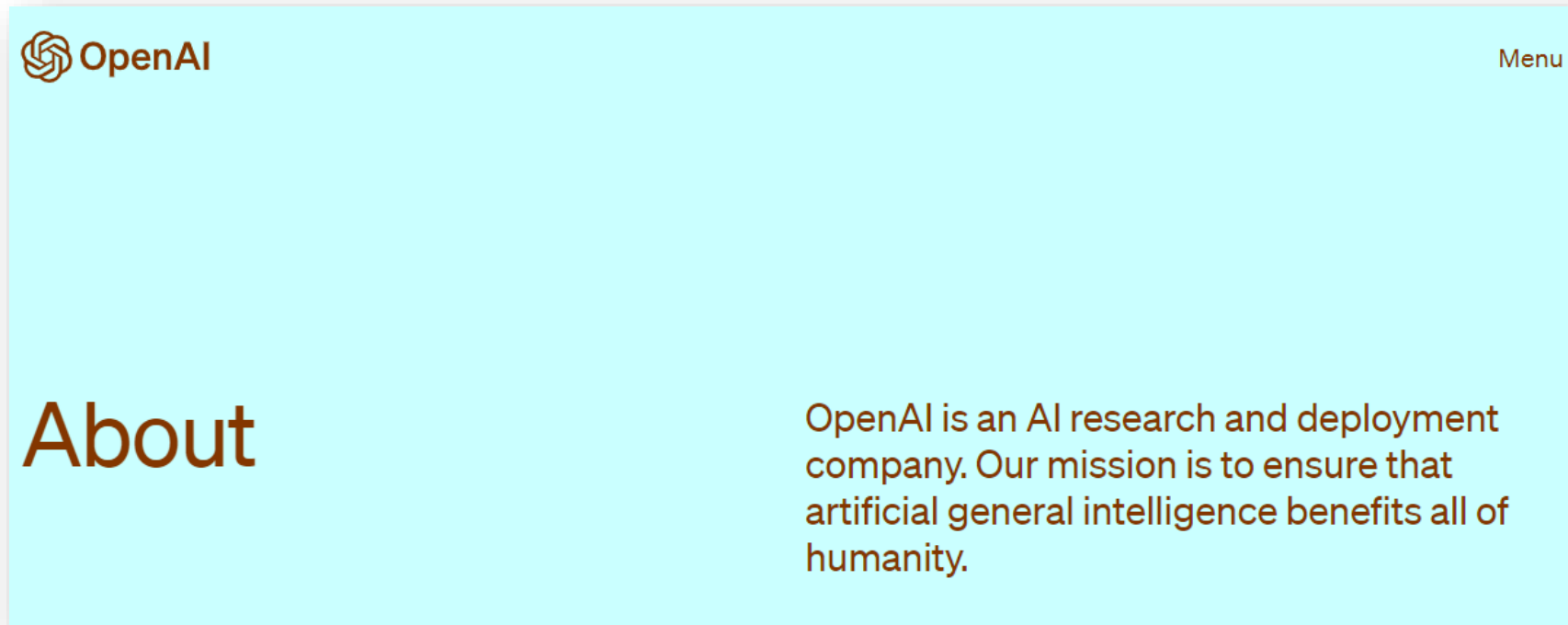
Systemic



Long term

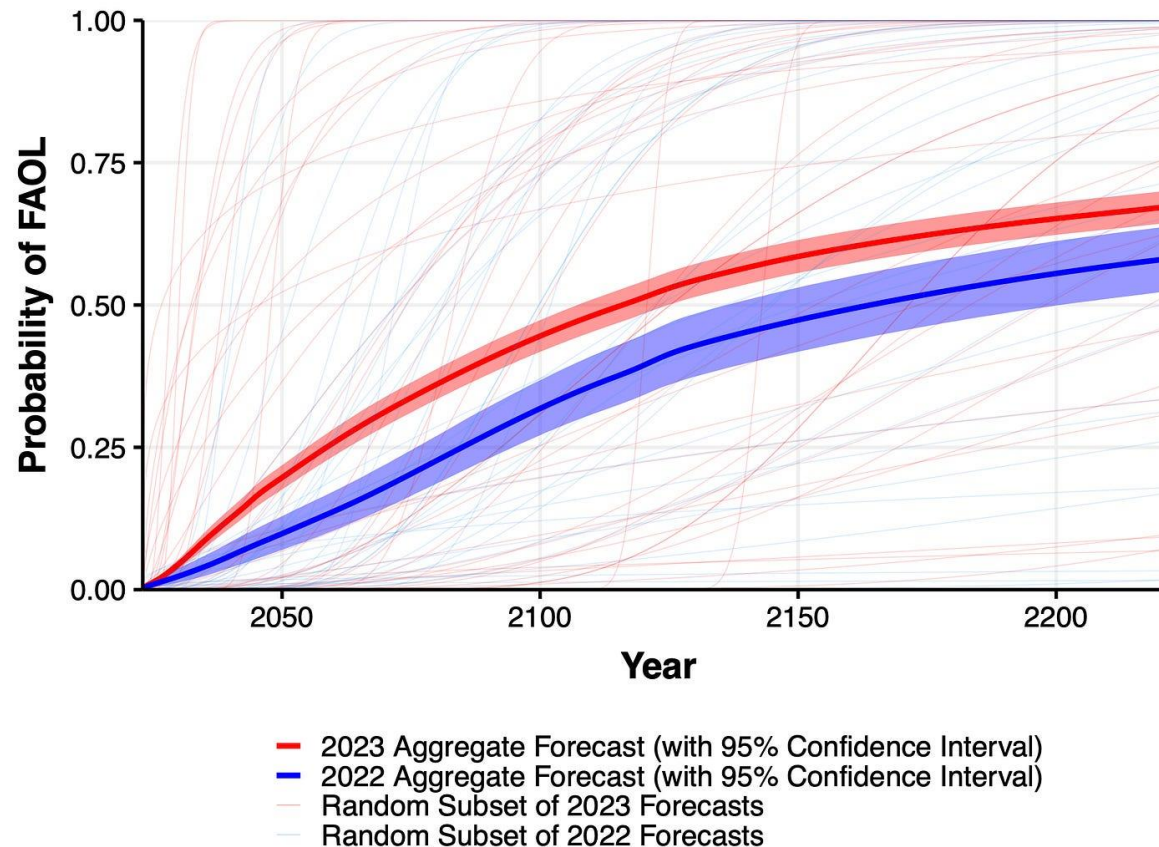
# Artificial General Intelligence (AGI)

- **AGI:** The hypothetical intelligence of a machine that has the capacity to understand or learn any intellectual task that a human being can.
- **AGI:** Like a median remote co-worker (Sam Altman)



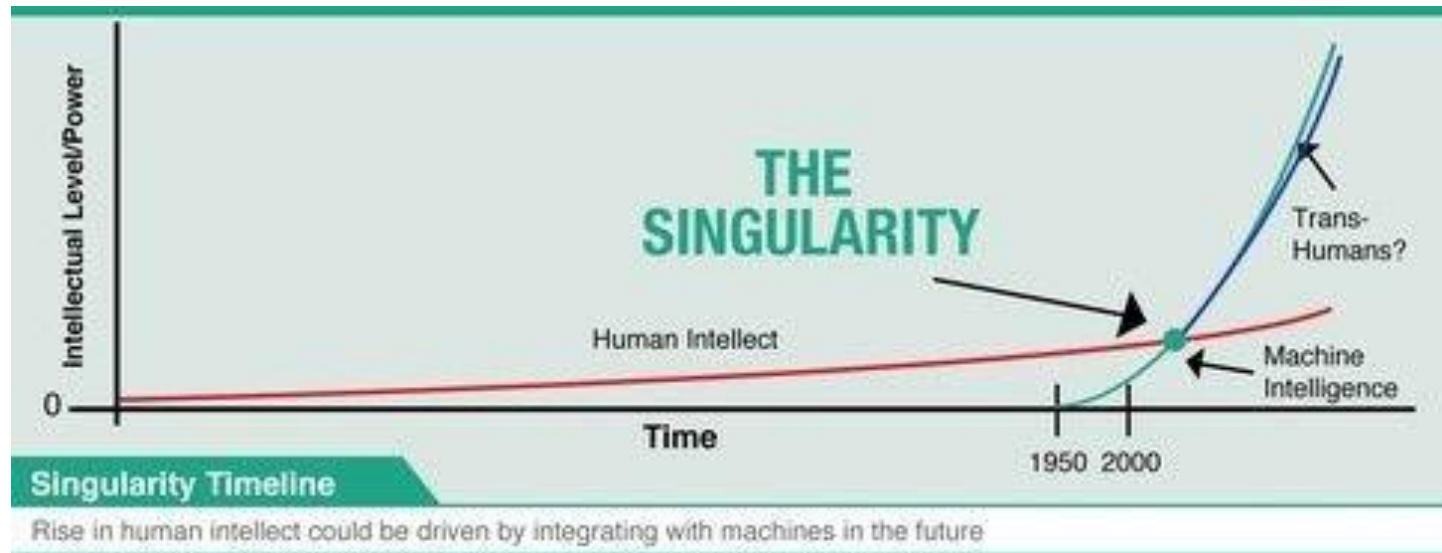
# Artificial General Intelligence (AGI)

- Survey among 2000+ AI experts: full automation of labor (FAOL)



# Technological Singularity

- Singularity is the point at which “technological growth becomes **uncontrollable** and irreversible, resulting in **unforeseeable changes** to **human civilization**”
- “Artificial Super Intelligence”



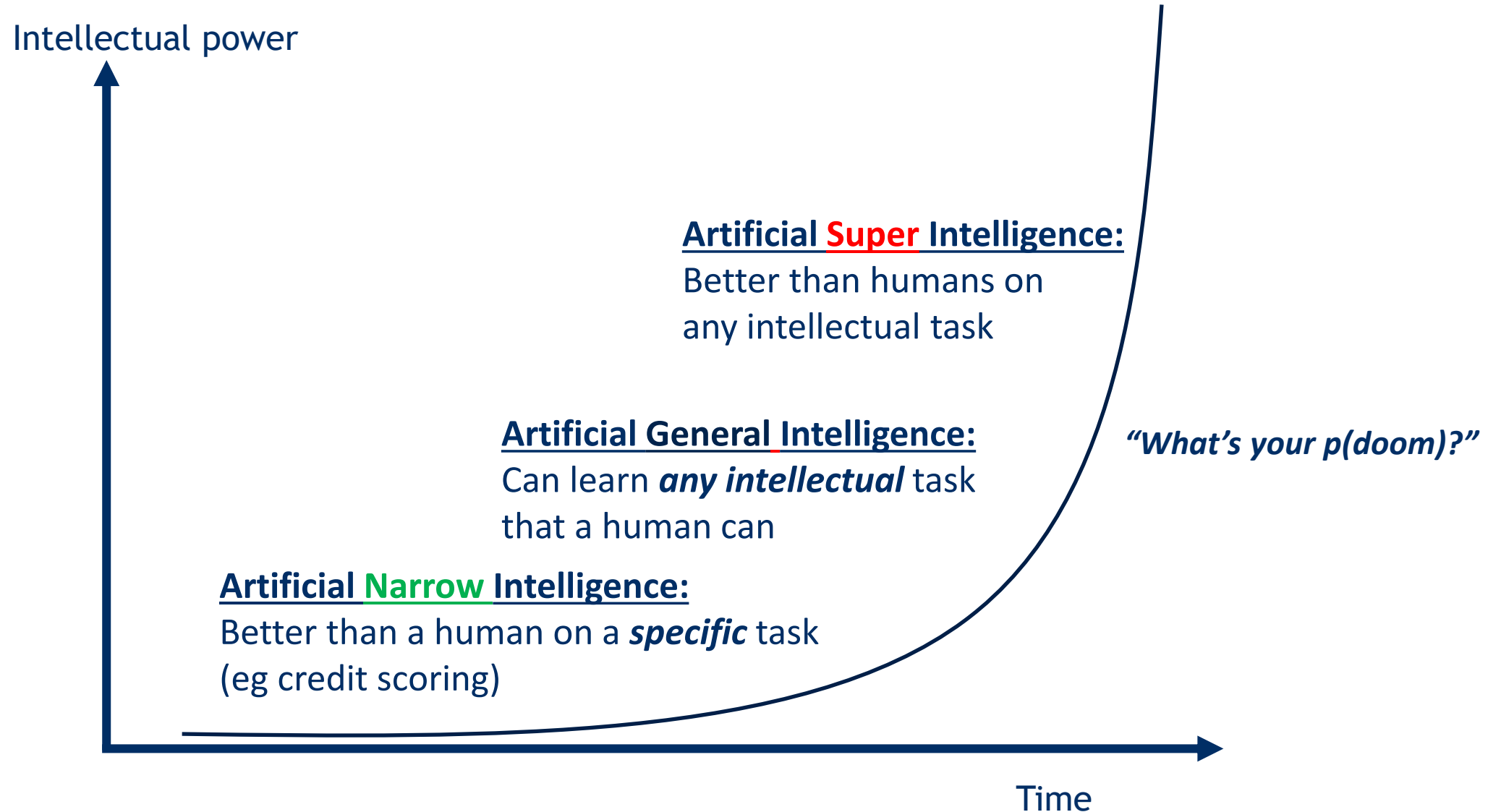
<https://innovationtorevolution.wordpress.com/2014/10/29/technological-singularity-from-fiction-to-reality/>

# Human Extinction

- $p(\text{doom})?$

- <0.01%** **Yann LeCun** one of three godfathers of AI, works at Meta (less likely than an asteroid)
- 10%** **Vitalik Buterin** Ethereum founder (Specifically means AI takeover)
- 10%** **Geoff Hinton** one of three godfathers of AI (wipe out humanity in the next 20 years)
- 9-19.4%** **Machine learning researchers** (From 2023, depending on the question design, median values: 5-10%)
- 15%** **Lina Khan** head of FTC
- 10-20%** **Paul Christiano** (Cumulative risks go to 50% when you get to human-level AI)
- 10-25%** **Dario Amodei** CEO of Anthropic
- 20%** **Yoshua Bengio** one of three godfathers of AI
- 20-30%** **Elon Musk** CEO of Tesla, SpaceX, X
- 5-50%** **Emmett Shear** Co-founder of Twitch, former short-term CEO of OpenAI
- 30%** **AI Safety Researchers** (Mean from 44 AI safety researchers in 2021)
- 33%** **Scott Alexander** Popular Internet blogger at Astral Codex Ten
- 35%** **Eli Lifland**

# Types of AI





# Conclusion

- **Embrace AI**
  - Age of AI
  - Fast moving
  - *“AI is not gonna take your job, someone who understands AI is going to take your job. **Get good at it.**”* Scott Galloway
- **Be aware of the risks**



# Questions



**Book:** David Martens

Data Science Ethics: Concepts, Techniques and Cautionary Tales  
*Oxford University Press* (2022)

272 pages

[www.dsethics.com](http://www.dsethics.com)

## Contact

**Research:** [david.martens@uantwerpen.be](mailto:david.martens@uantwerpen.be)

**Advisory:** [david.martens@searchingpi.com](mailto:david.martens@searchingpi.com)