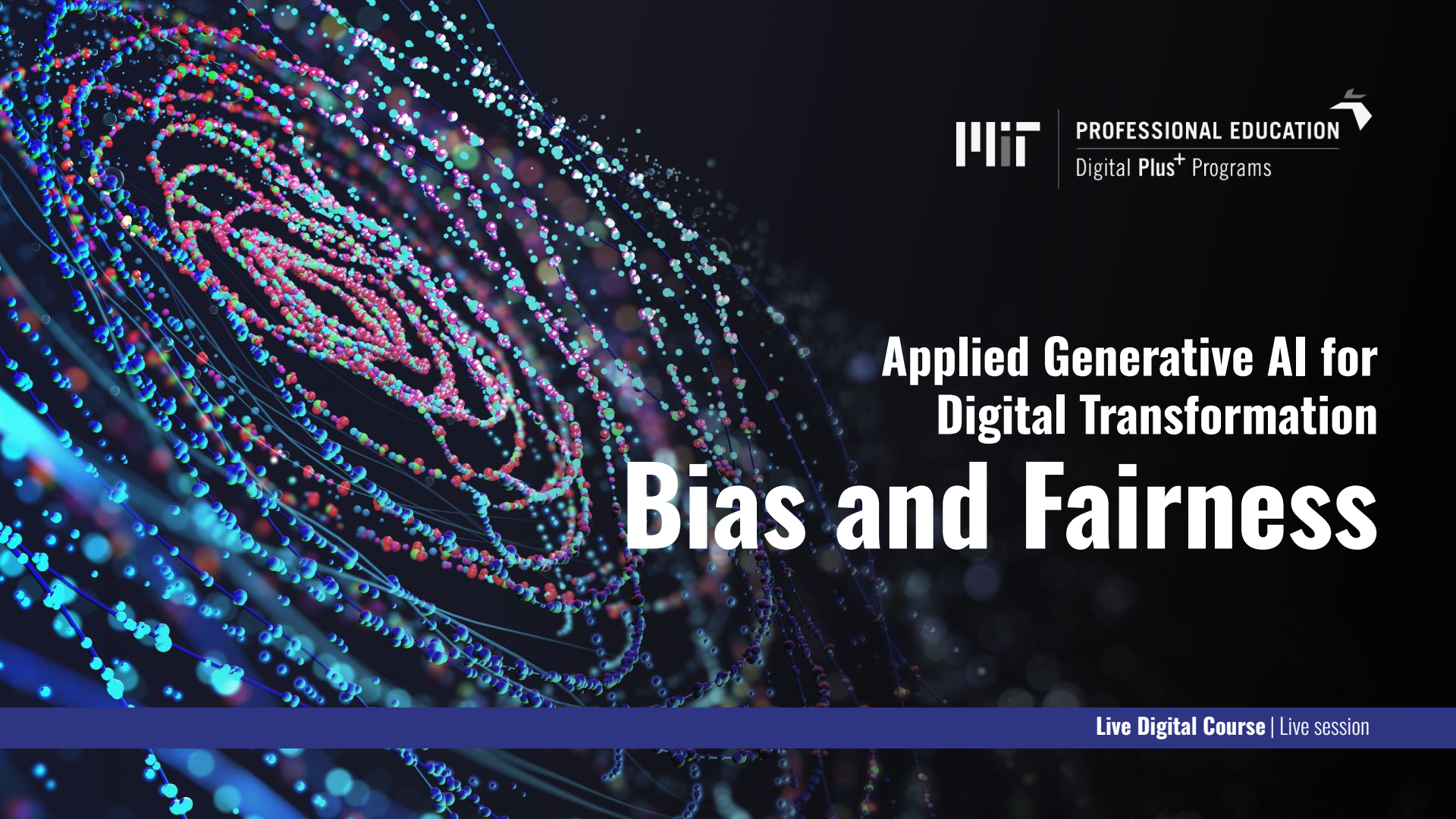
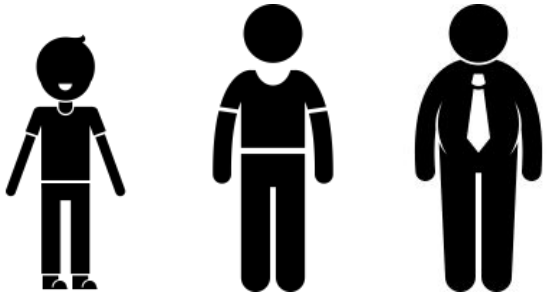

PROFESSIONAL EDUCATION
Digital Plus+ Programs
Applied Generative AI for Digital Transformation
Bias and Fairness
Live Digital Course | Live session

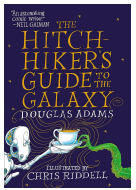
Common reactions to technology

1. Anything that is in the world when you’re born is normal and ordinary and is just a natural part of the way the world works.

2. Anything that's invented between when you’re fifteen and thirty-five is new and exciting and revolutionary and you can probably get a career in it.

3. Anything invented after you're thirty-five is against the natural order of things.





Reactions described by Douglas Adams

Poll: Can only human creators win a Grammy or an Oscar?

- Yes
- No

https://commons.wikimedia.org/wiki/File:GuitareClassique5.png

Nothing is so painful to the human mind as a great and sudden change

Mary Shelley 1818

Technology fears: we do not do well with change

When radio was introduced:

- Educators: radio threatens culture of U.S.
- Radio blamed for weather changes, business decline
- Radio harms children, poor grades
- ⅓ of U.K. children cannot read properly
- Radio will kill churce attendance
- No politics on radio. Every speech must be submitted to authorities first.
- Children robbed of values, spikes in crime.
- Balance of elections tipped

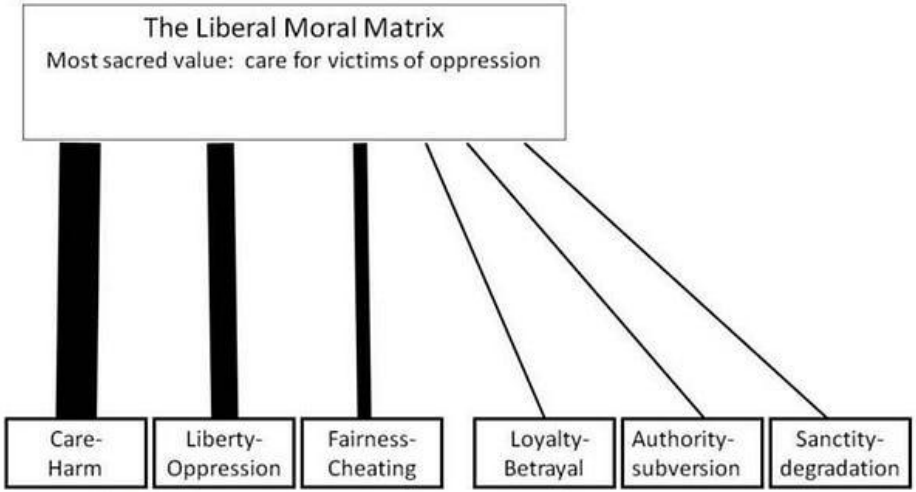


Figure 12.2. *The moral matrix of American liberals.*

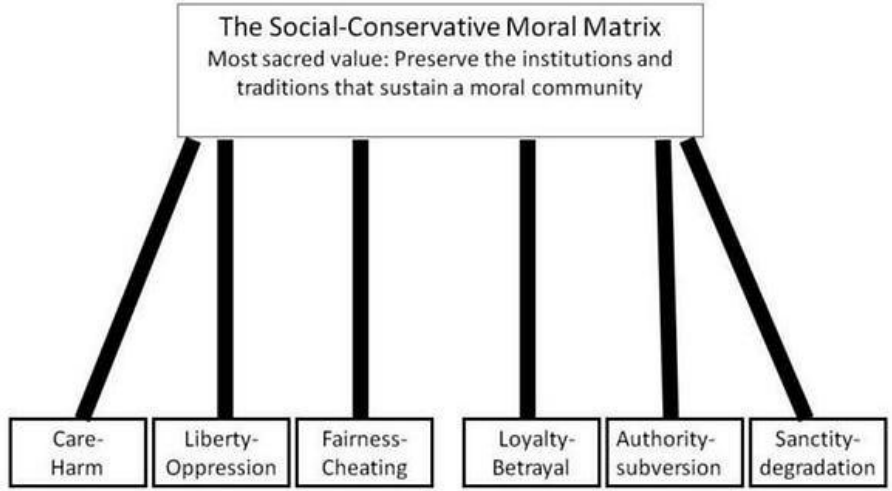


Figure 12.4. *The social conservative moral matrix.*

Haidt, Jonathan. The righteous mind: Why good people are divided by politics and religion. Vintage, 2012.

AI Boomers vs Doomers

The Doomer Advantage

- Most new ideas do not work
- Same as new business ideas
- It’s easy to critizise, you are right most of the time
- It’s why new ideas leave organizations (e.g. transformer)
- New ideas do not always fail

The Boomer Disadvantage

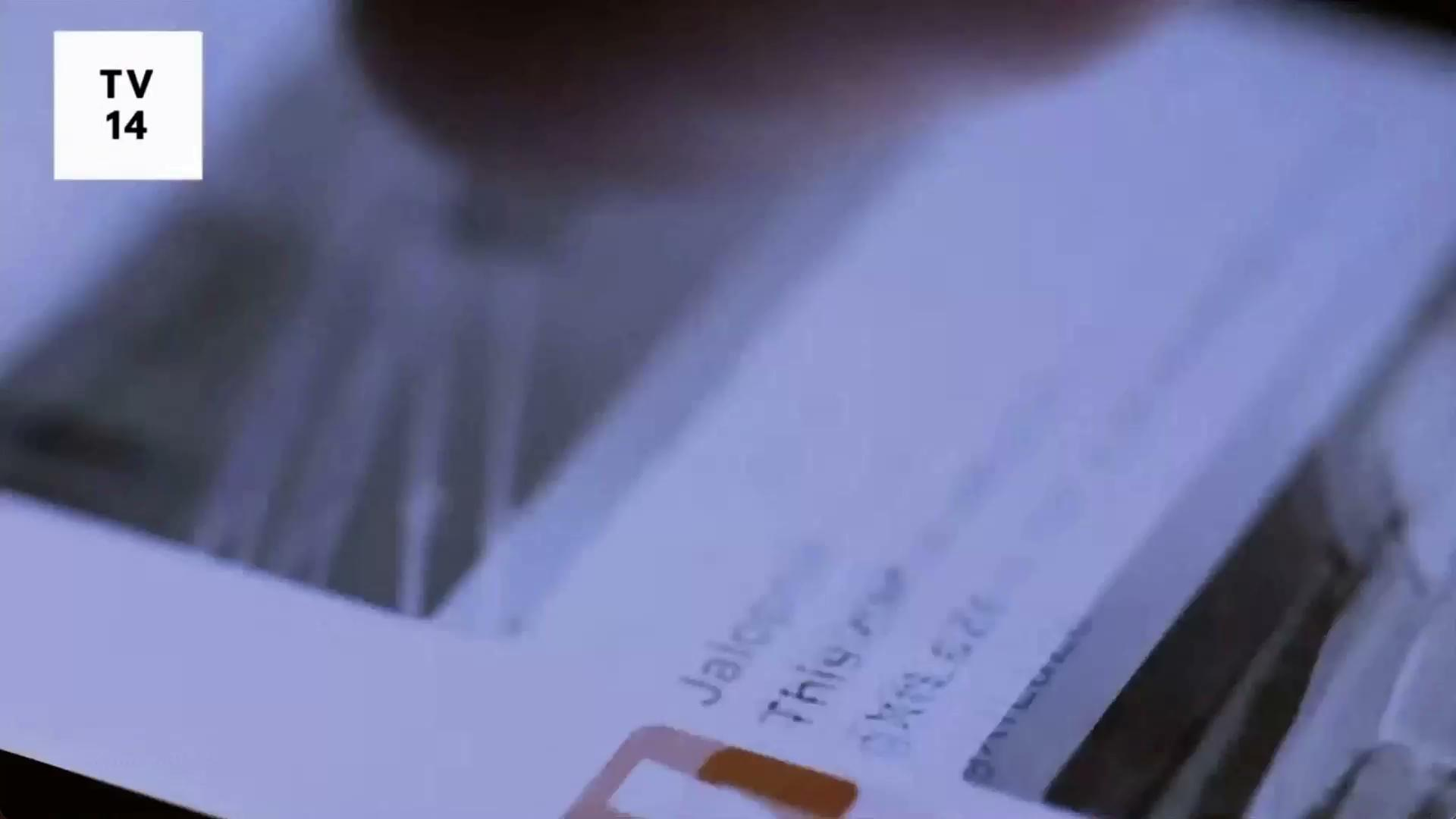
- High risk
- Industry average, 1-of-8
- BP, 1-of-5
- Balance of risk/reward, business/data, 2-of-3
- Addresing risk:
  - DevOps
  - The Lean Startup by Eric Ries
  - Ask Your Developer by Jeff Lawson
  - Team of Teams by General Stanley McChrystal

Tech has rarely stifled human creativity

- Electronic synthesizers did not eliminate the need for people to play music instruments
- Auto-Tune didn’t make singing on pitch obsolete
- Photography didn’t kill painting
- Photography digitization did not eliminate professional photographers
- Vinyl records outsell CDs for the second year running

Video - AI Ethics

https://bit.ly/3Euyuo2

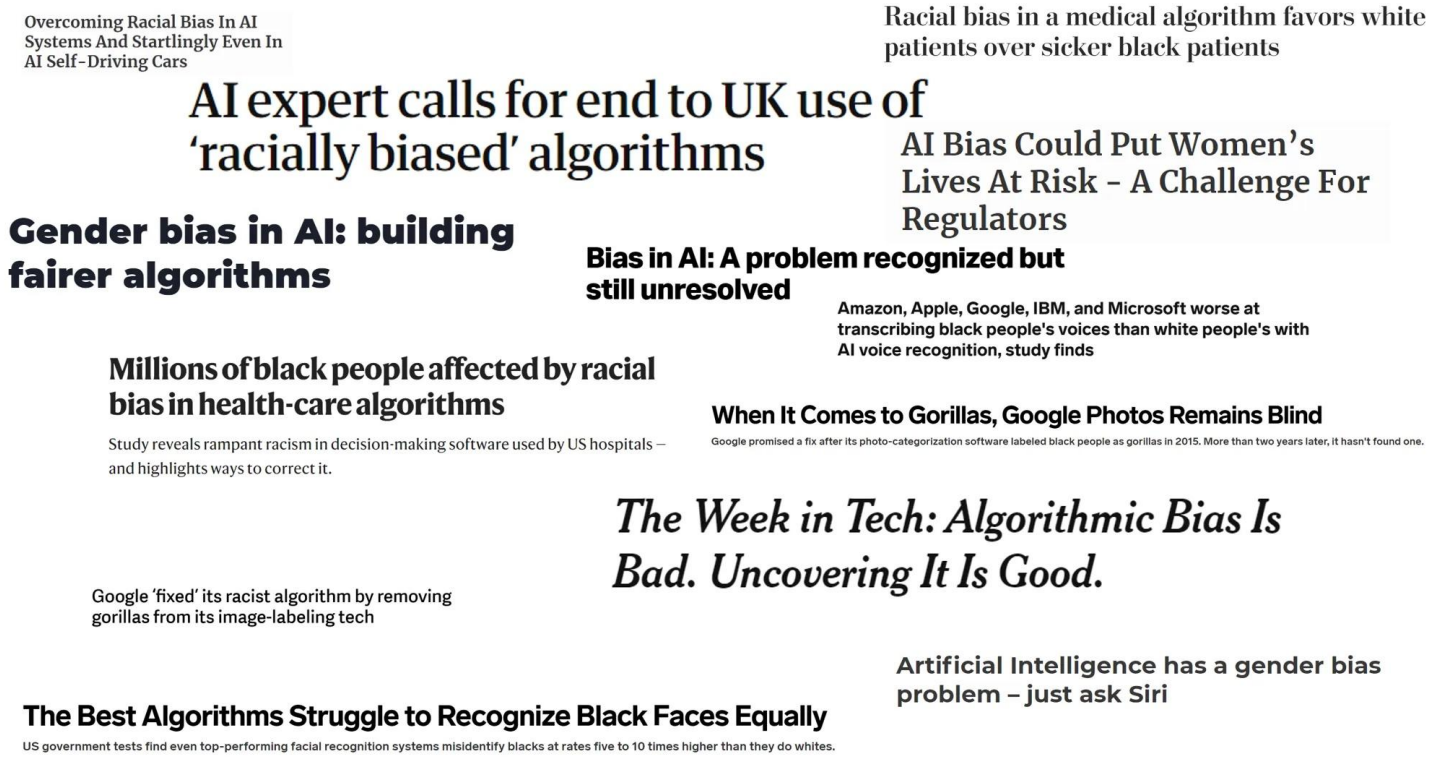

TV
14

Facial recognition algorithm

In 2020, Detroit police arrested a Black man for shoplifting almost $4,000 worth of watches from an upscale boutique. He was handcuffed in front of his family and spent a night in lockup. After some questioning, however, it became clear that they had the wrong man. So why did they arrest him in the first place?

The reason: a facial recognition algorithm had **matched the photo on his driver's license to grainy security camera footage**.

Headlines

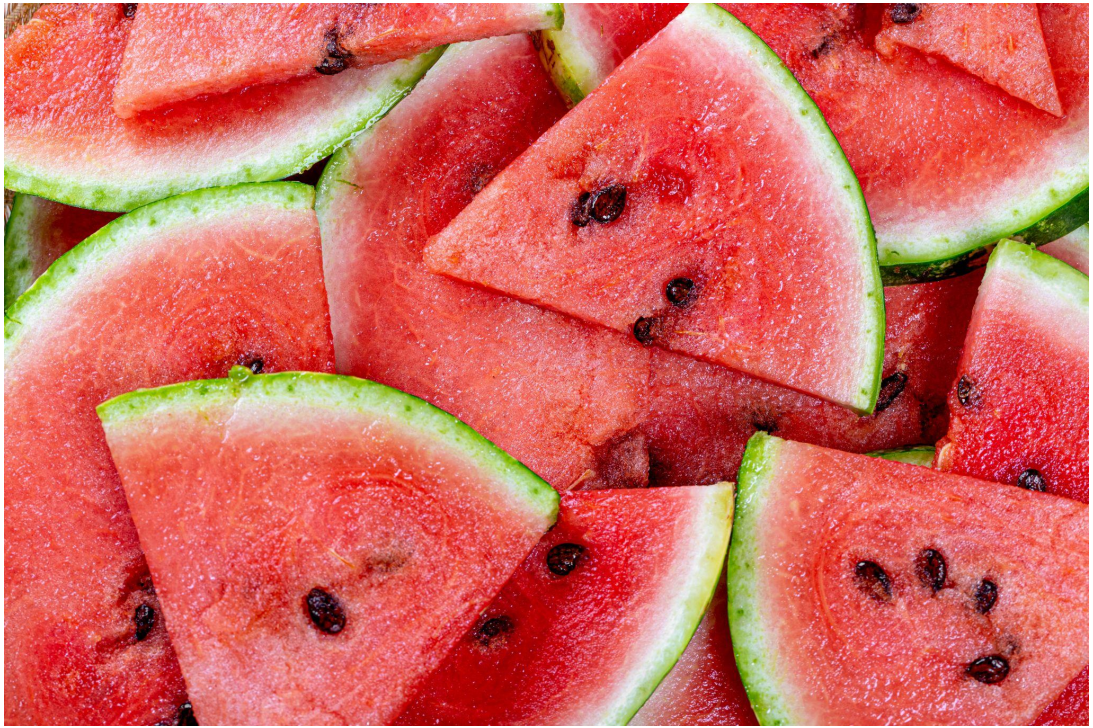


https://towardsdatascience.com/algorithm-bias-in-artificial-intelligence-needs-to-be-discussed-and-addressed-8d369d675a70

What is bias?

How would you describe it?

- Watermelon
- Watermelon slices
- Watermelon seeds
- Juicy watermelon



How many of you though to describe it as red watermelon?

- Yellow Watermelon
- Yellow Watermelon slices
- Yellow Watermelon seeds
- Juicy Yellow watermelon

Why did we not say "red" watermelon in the first image?

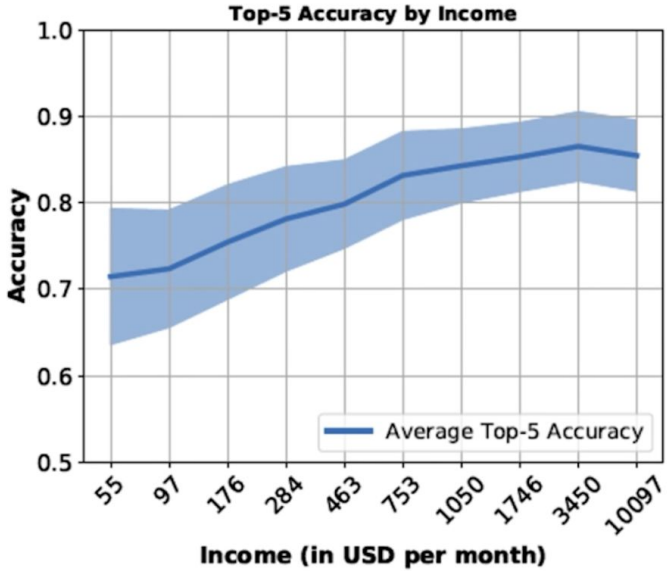
- We did not think of the image as red watermelon
- Red is the expected color for a watermelon
  - It is our geographical bias
  - It would be different in other parts of the world
- Biases have always existed, what is the big deal?
- Our previous bias was localized and geographically confined
- In our globally connected world, an AI bias has global reach

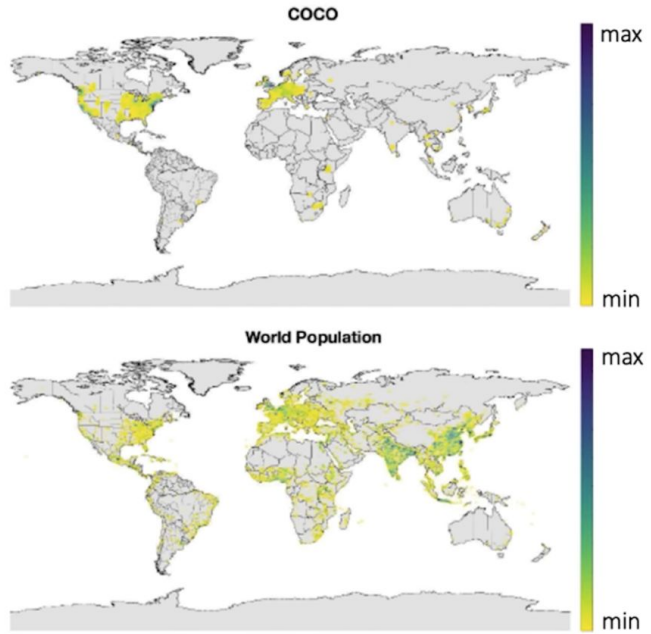
Bias in facial detection

Independent Study I

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Bias correlation with income and geography

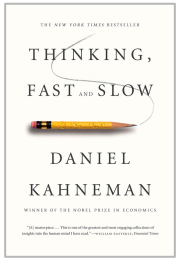




De Vries, et al. "Does object recognition work for everyone?.", 2019

Labelling challenge

- We label and categorize the world to reduce complex sensory inputs into simplified groups that are easier to work with.
- There is going to be an "expected" representation, a typical representation - Thinking, Fast and Slow, Daniel Kahneman.
- Biases can arise when particular labels confound decisions (mixup or confusion, specially against expectation) - human or artificial.



Bias in the AI life cycle

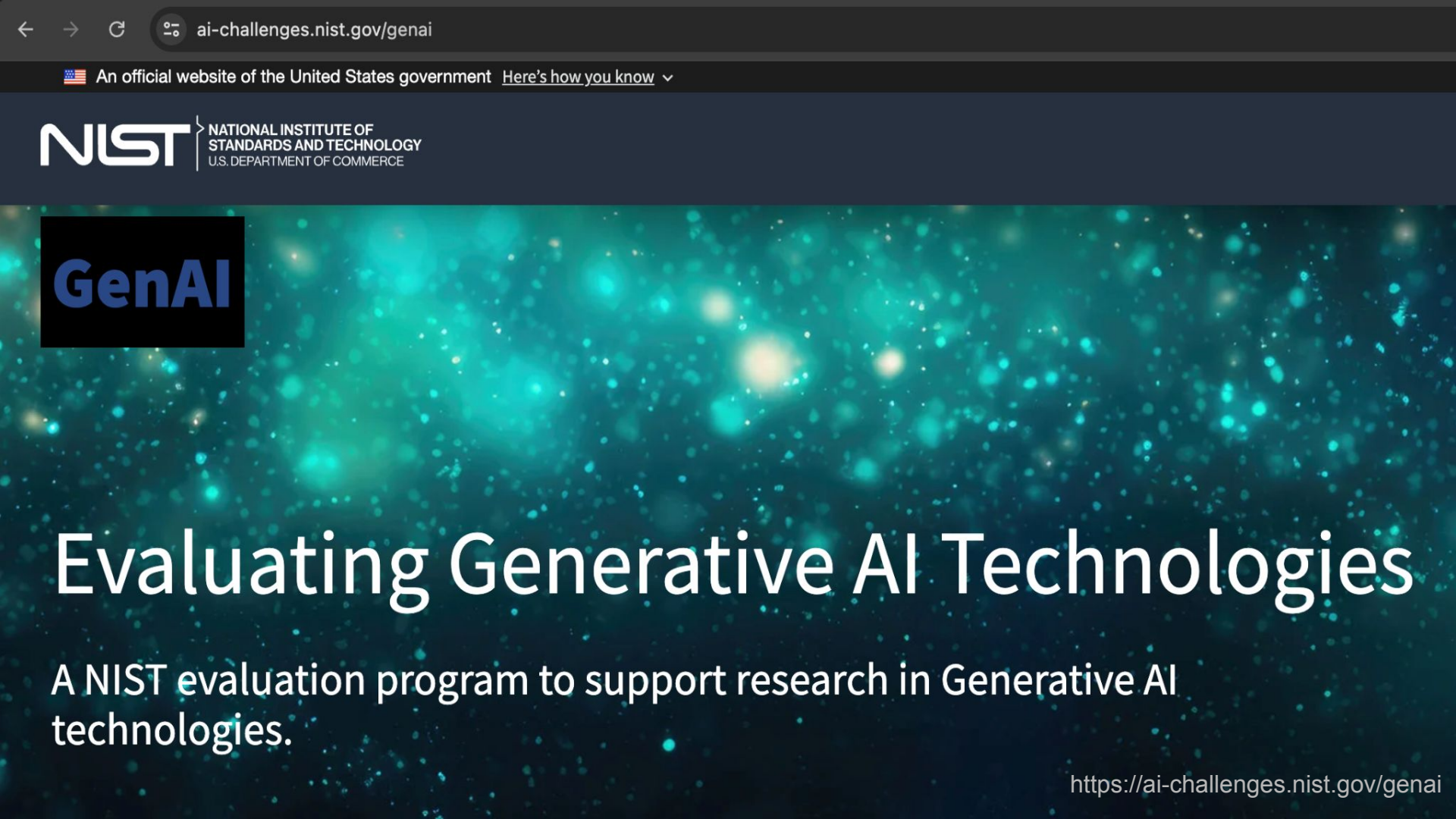
- **Data:** imbalances with respect to class labels, features, inputs
- **Model:** interpretability, and performance metrics
- **Training and deployment:** feedback loops that perpetuate biases
- **Evaluation:** data subgroups
- **Interpretation:** human errors and biases distort meaning of results

Voices & Regulation

Fears grow over the potential use of AI

- Pause Giant Ai Experiments: An Open Letter, Future Life
  https://futureoflife.org/open-letter/pause-giant-ai-experiments/
- Dangers of Stochastic Parrots, ACM, 1375 citations
  https://dl.acm.org/doi/pdf/10.1145/3442188.3445922
- AI Accountability Policy Request for Comment, US
  https://ntia.gov/issues/artificial-intelligence/request-for-comments
- Why AI Will Save the World by Marc Andreessen
  https://a16z.com/2023/06/06/ai-will-save-the-world/
- Adaptative Ethics for Digital Transformation, Mark Schwartz
  https://www.amazon.com/dp/1950508714
- UK's pro-innovation AI regulation, UK

  https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1146542/a_pro-innovation_approach_to_AI_regulation.pdf


ai-challenges.nist.gov/genai
An official website of the United States government Here's how you know
NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY U.S. DEPARTMENT OF COMMERCE
GenAI
Evaluating Generative AI Technologies
A NIST evaluation program to support research in Generative AI technologies.
https://ai-challenges.nist.gov/genai

AISI AI SAFETY INSTITUTE

Inspect

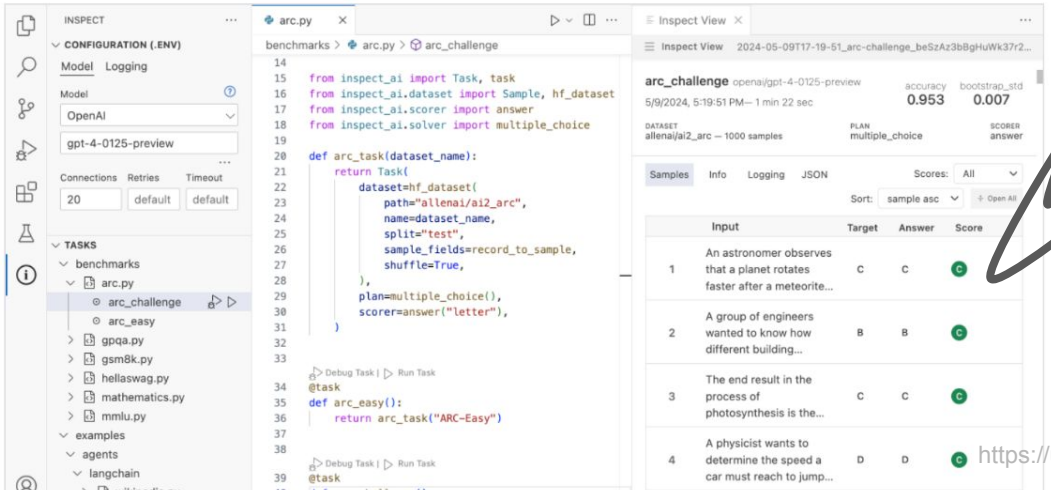
Inspect

An open-source framework for large language model evaluations

Welcome

Welcome to Inspect, a framework for large language model evaluations created by the UK AI Safety Institute.

Inspect provides many built-in components, including facilities for prompt engineering, tool usage, multi-turn dialog, and model graded evaluations. Extensions to Inspect (e.g. to support new elicitation and scoring techniques) can be provided by other Python packages.



Test / Code

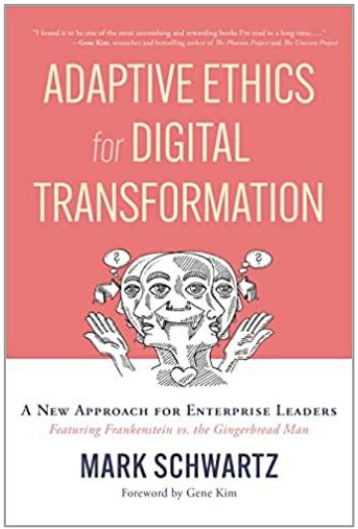
Report an issue

Safety, security, trust

- Safety
  - 1) Commit to internal and external red-teaming of models
  - 2) Work toward information sharing among companies and governments
- Security
  - 3) Invest in cybersecurity and insider threat safeguards
  - 4) Incent third-party discovery and reporting
- Trust
  - 5) Develop mechanisms that detection of AI-generated content
  - 6) Publicly report model or system capabilities
  - 7) Prioritize research on societal risks
  - 8) Develop AI systems to address society’s greatest challenges

https://openai.com/index/moving-ai-governance-forward/

Adaptive Ethics for Digital Transformation



Link to the conversation/interview
https://bit.ly/3CRqgFQ

Control

Traditional AI

Human created AI

(autonomous driving)

Traditional AI

Human created & guided AI

(image recognition - automated tolls)

Generative AI

Robot advisor

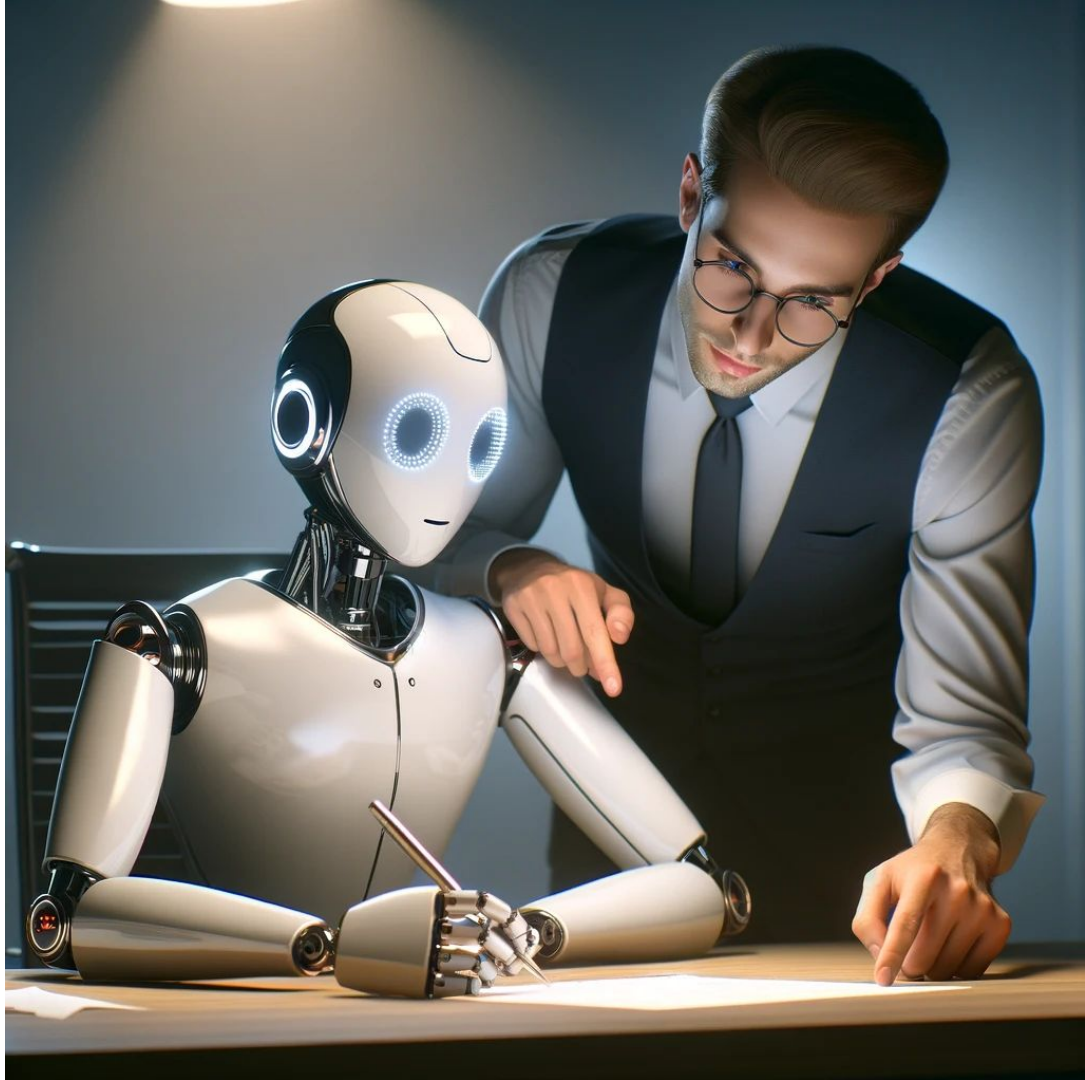
(domain expert + LLM, e.g. AI code advisor, human-AI collaboration)

Scale up

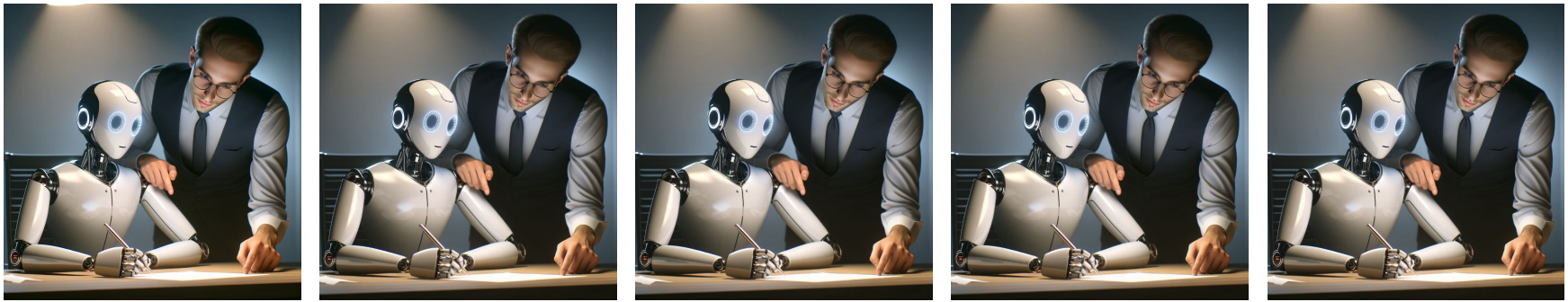
Generative AI

Robot creator, human reviewer

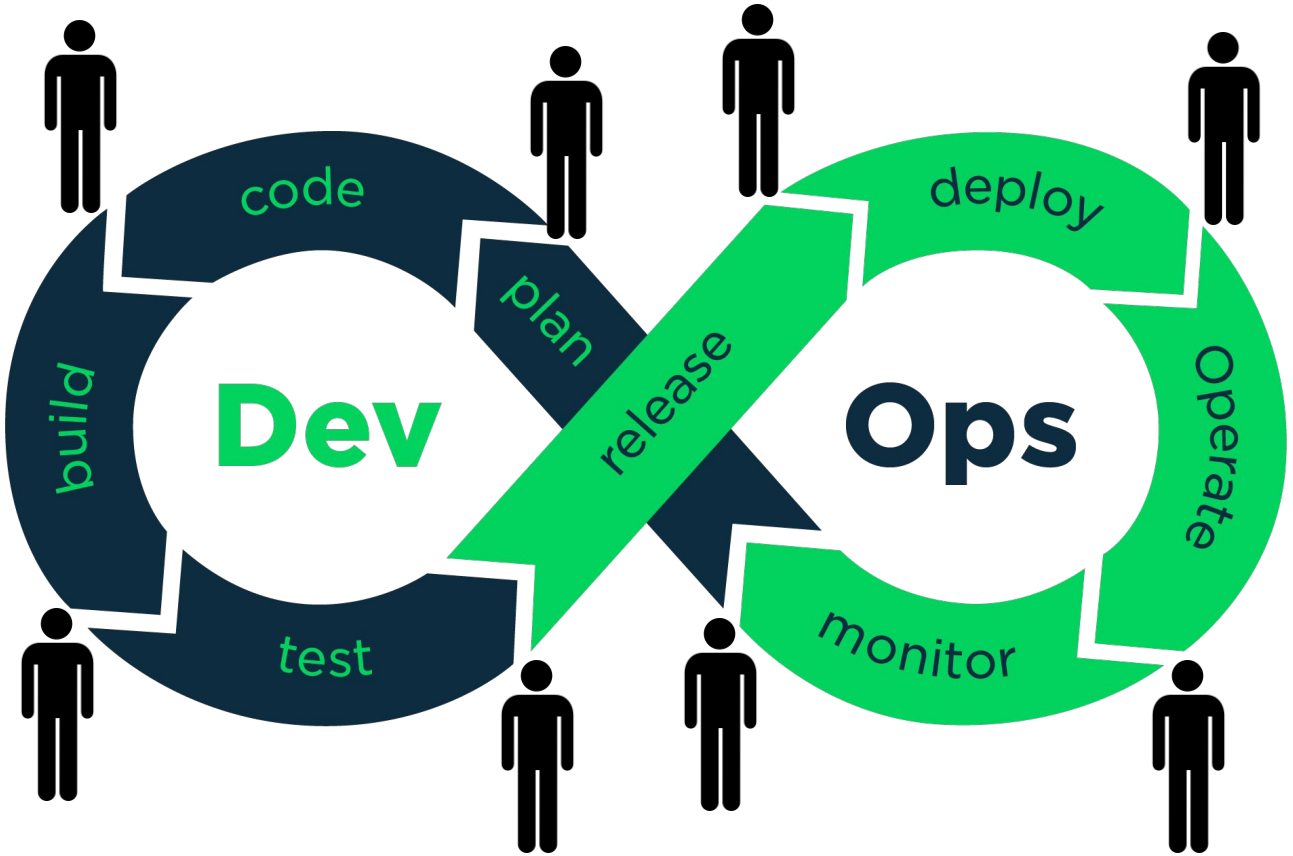
(domain expert + LLM, automated marketing campaign)



Scale up



Human Quality Control




Currently humans look after software systems.
Sometimes they are unpredictable.
Waze &
Canary
Story

Generative AI

Scaling Leadership
Human orchestrator

Generative AI

Robots leading robots

Generative AI

Scaling Leadership
Robot orchestrator

Scalability for sure, what about control?

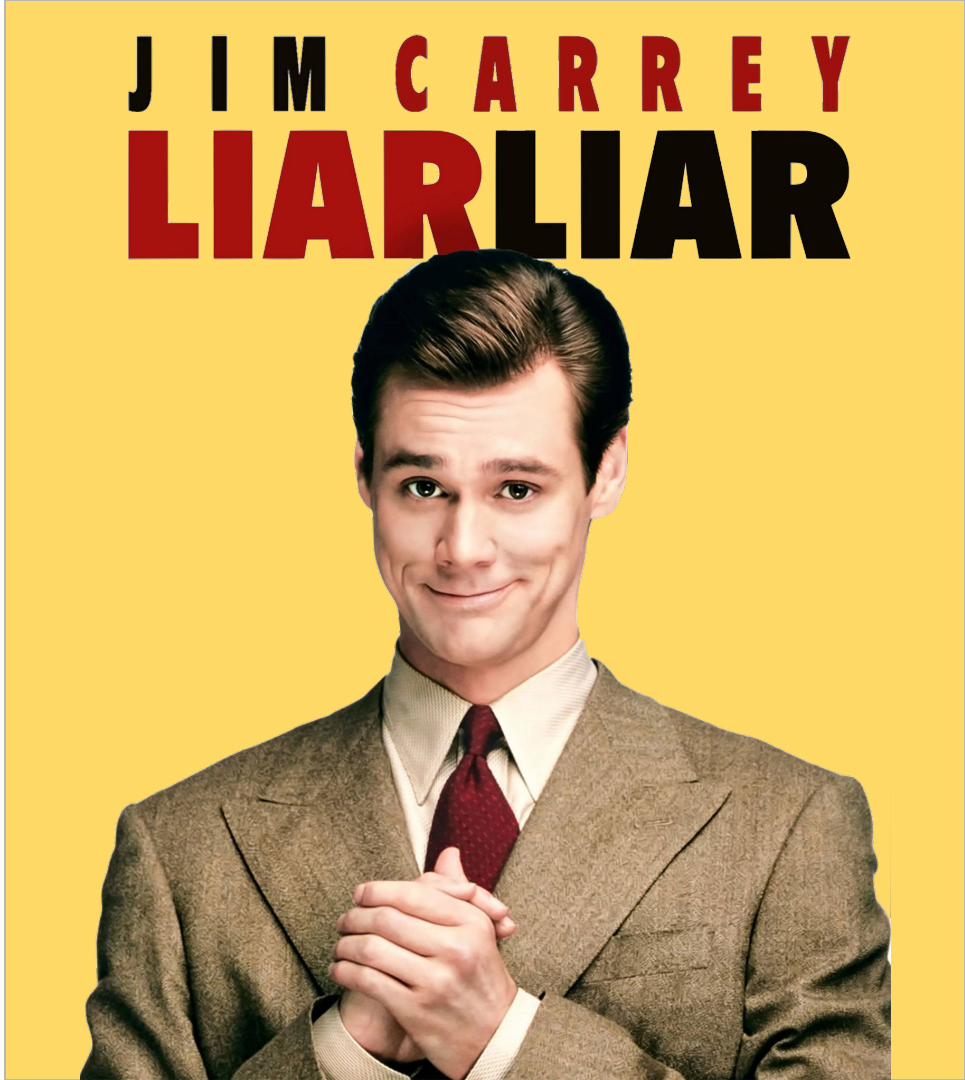
We already have automated systems

Human supervising humans

Human systems also fail

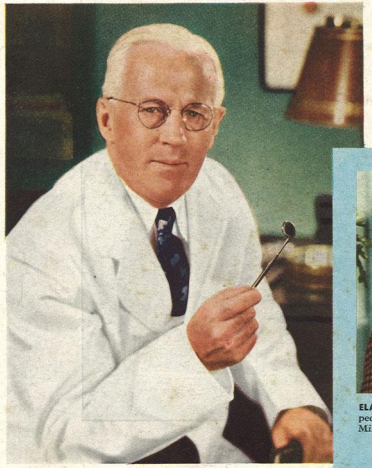
How well do we handle misinformation?

Humans have been lying for a long time



Noted throat specialists report on 30-day test of Camel smokers...

NOT ONE SINGLE CASE OF THROAT IRRITATION *due to smoking* CAMELS!



Yes, these were the findings of noted throat specialists after a total of 2,470 weekly examinations of the throats of hundreds of men and women who smoked Camels—and only Camels—for 30 consecutive days.

ELANA O'BRIAN, real estate broker, one of the hundreds of people from coast to coast who made the 30-Day Test of Camel Mildness under the observation of noted throat specialists.

...AND THOUSANDS MORE AGREE!

"CAMELS AGREE with my throat—and they sure taste great!" says Ed Paxton, chemical engineer.

EDITORIAL ASSISTANT Virginia Walcutt: "Camels met the test—they certainly agree with my throat!"

MICHAEL DOUGLAS, singer: "Camels give me the kind of smoke I like—lots of flavor and plenty mild!"

MISS LEE TELLER, secretary: "The cigarette that really agrees with my throat—Camels!"

"THE 30-DAY TEST taught me there's no cigarette like a Camel!" Tod Crone, travel agency owner.

SPORTSWOMAN Jean French: "Camels tasted so good I've changed to Camels for keeps!"

Start your own 30-Day Camel MILDNESS Test Today!

It's fun! All you do is smoke Camels for 30 days. Compare them in your "T-Zone" (T for taste, T for throat). See if that rich, full Camel flavor and that cool Camel mildness doesn't win you to Camels for keeps.

R. J. Reynolds Tobacco Co., Winston-Salem, N. C.

How mild can a cigarette be?

In a recent coast-to-coast test, hundreds of men and women smoked Camels—and only Camels—for 30 consecutive days. These people smoked on the average of one to two packs a day. Each week, during the entire test period, throat specialists examined these Camel smokers. A total of 2470 careful examinations were made. The doctors who made the throat examinations of these Camel smokers reported:

"NOT ONE SINGLE CASE OF THROAT IRRITATION due to smoking CAMELS!"

According to a Nationwide survey:

MORE DOCTORS SMOKE CAMELS

than any other cigarette

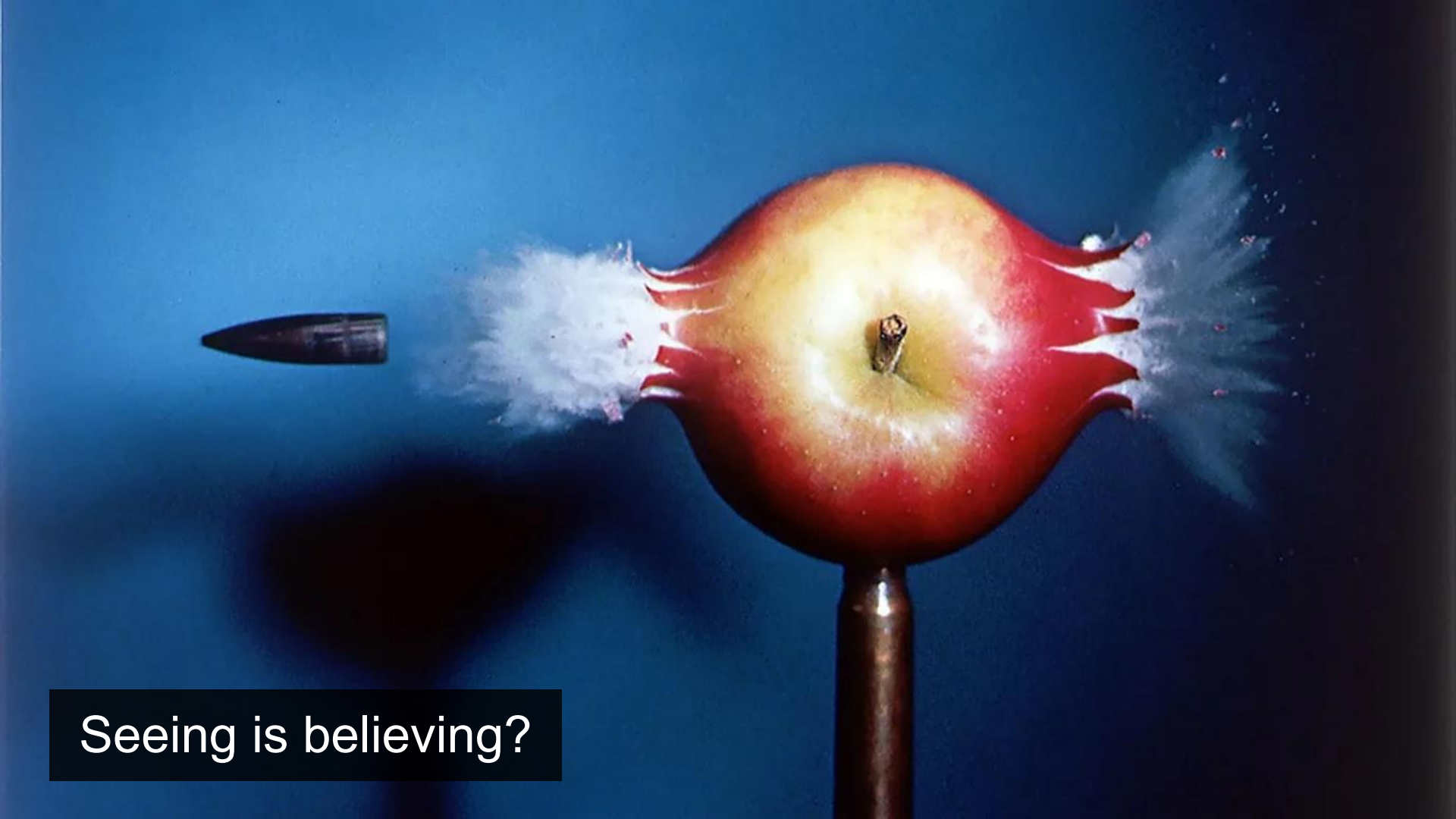
Doctors smoke for pleasure, too! And when three leading independent research organizations asked 113,597 doctors what cigarette they smoked, the brand named most was Camel!

Money-Back Guarantee!

Smoke Camels and test them in your own "T-Zone"—T for taste, T for throat. If, at any time, you are not convinced that Camels are the mildest cigarette you have ever smoked, return the package with the unused Camels and we will refund its full purchase price, plus postage. *(Signed)* R. J. Reynolds Tobacco Company, Winston-Salem, North Carolina.

We are used to being lied to

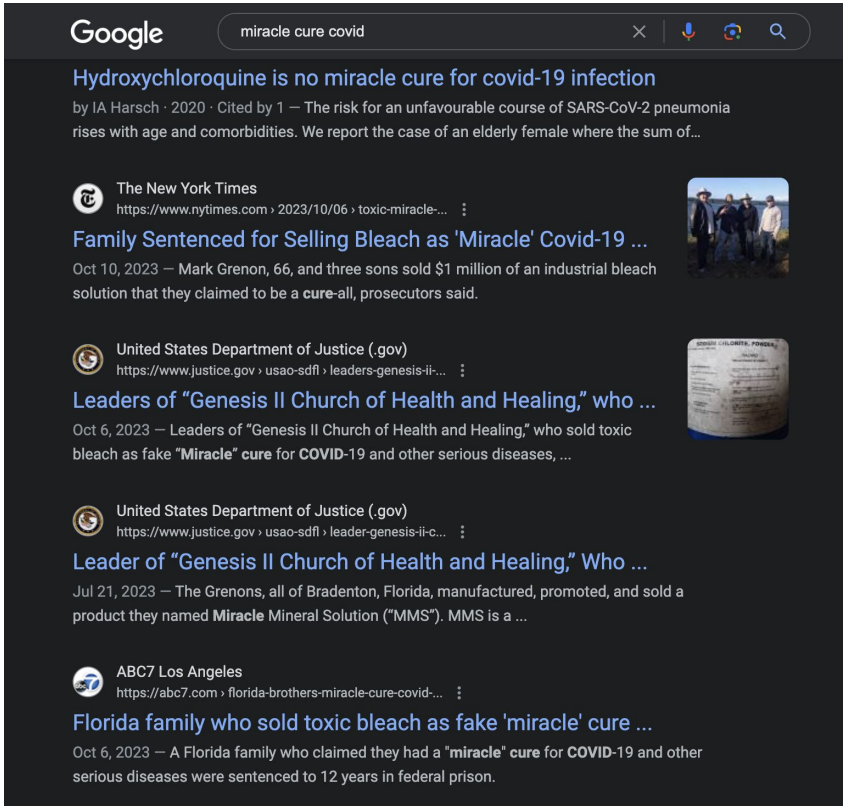
- It happens frequently
- Buying a car
- Web search
- Sales person
- We even expected it in certain contexts


Seeing is believing?

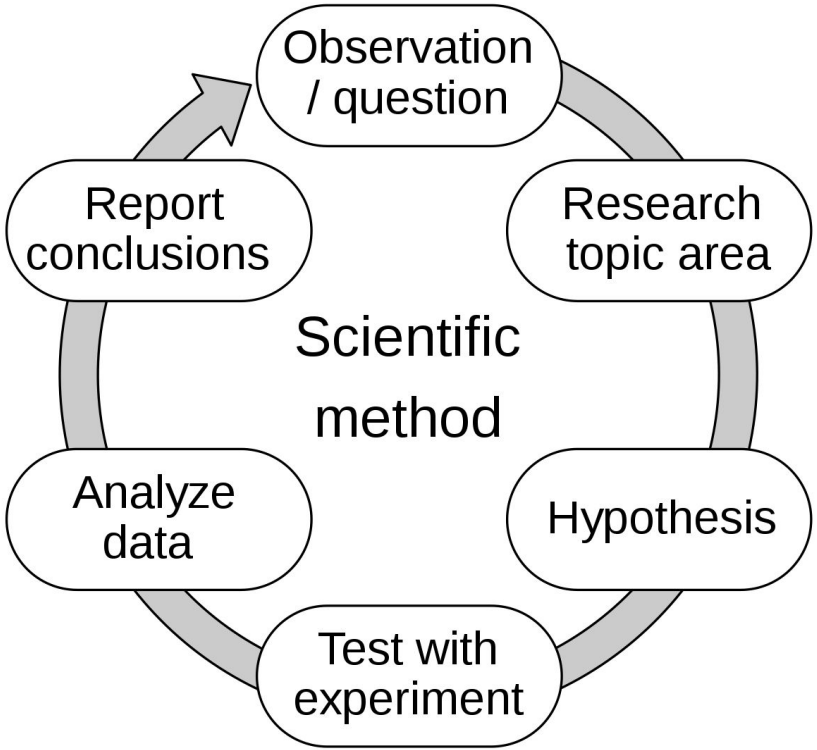
We already see and experience illusions


Film that is hyperrealistic

Pandemic misinformation - new threat or more visible threat? Government responsibility?



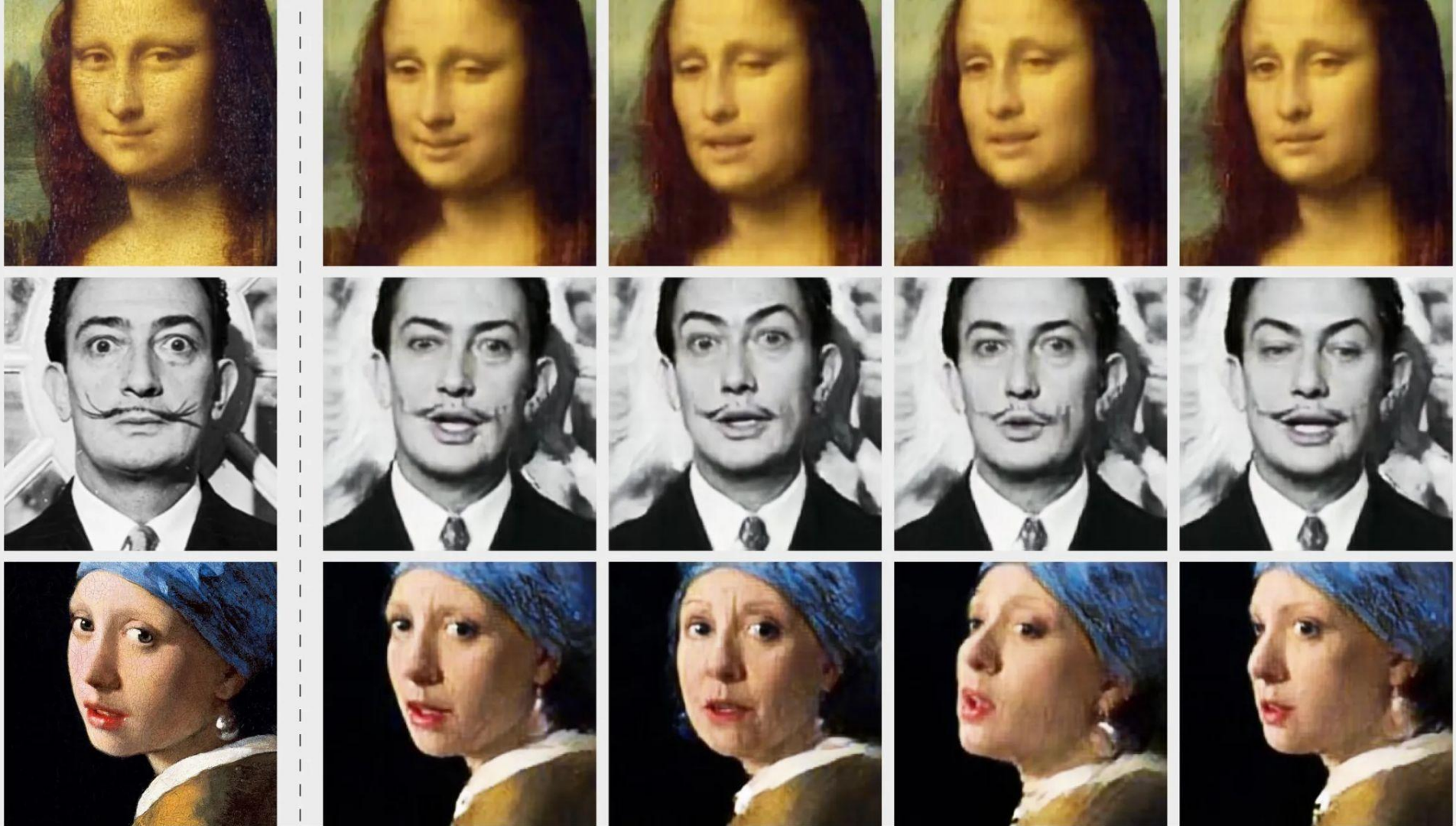
How does the scientific community determine what is true?



Can the science community be tricked?

- 100 psychology experiments repeated, less than half successful (Science 08/28/2015)
- 67 Economics papers 50% could no be independently replicated (Federal Reserve 9/4/2015)
- 53 'landmark' cancer publications, majority cannot be replicated (Nature, 3/28/2012)

How is AI the same/different? (exercise)



What do we object to?

- The cost reduction
- The speed
- The access
- The quality
- Why is different?
- Considerations for regulation

Cybersecurity Inspiration

Cybersecurity inspiration

- Red Teams
- External audits
- Practice recovery
- Practice press communications

Cybersecurity inspiration - NIST Framework

Capability	Description
Identify	What processes and assets need protection?
Protect	Implement appropriate safeguards to ensure protection of the enterprise's assets
Detect	Implement appropriate mechanisms to identify the occurrence of cybersecurity incidents
Respond	Develop techniques to contain the impacts of cybersecurity events
Recover	Implement the appropriate processes to restore capabilities and services impaired due to cybersecurity events

A/B test with human decision makers

- Monitor data - part of your data strategy
- Monitor models
- Anomalies are easy, logic problems are hard

Consider Open-sourcing your code, your data

- Like in security, more eyes helps
- Higher accountability
- Better code
- Model marketplace
- Model platform

Emerging titles

- Bias officer
- Algorithmic auditor
- Data Detective
- Data diversity officer
- VR X

The usual recommendations

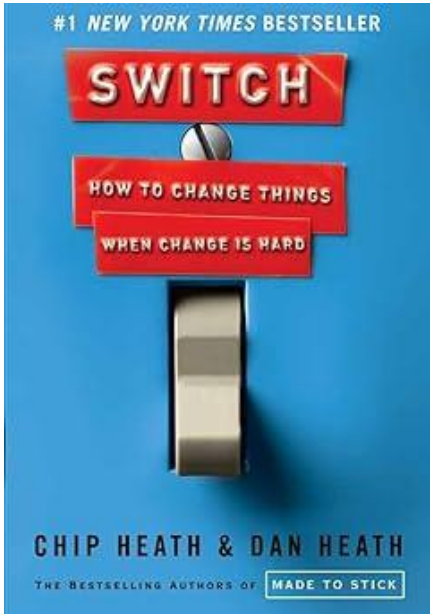
- Diversify the AI community
- Education
- Human-in-the-loop

Final Thoughts

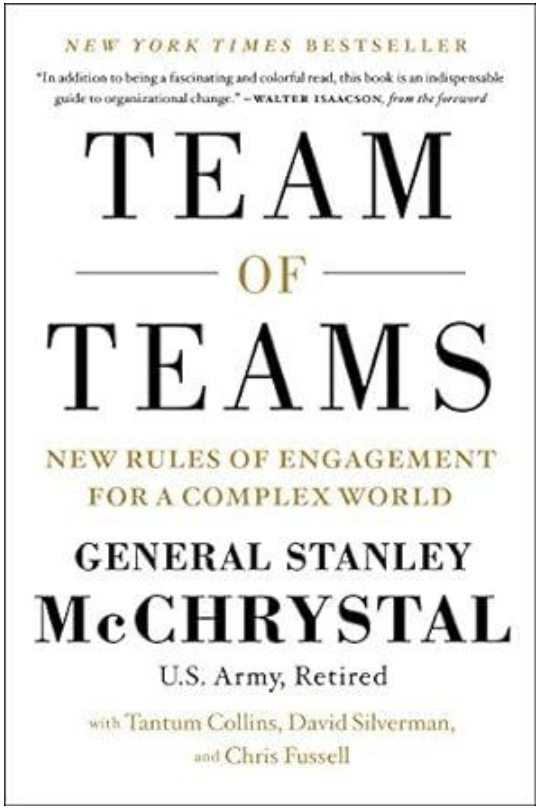
What is easier to change?

- Humans
- Machines

A great ethics document is not enough - Enron had one of the best. You need change management.



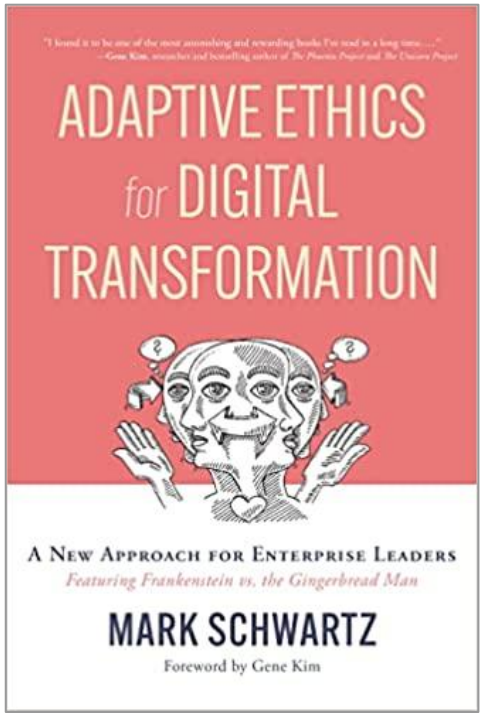
Managing change in a fast changing world

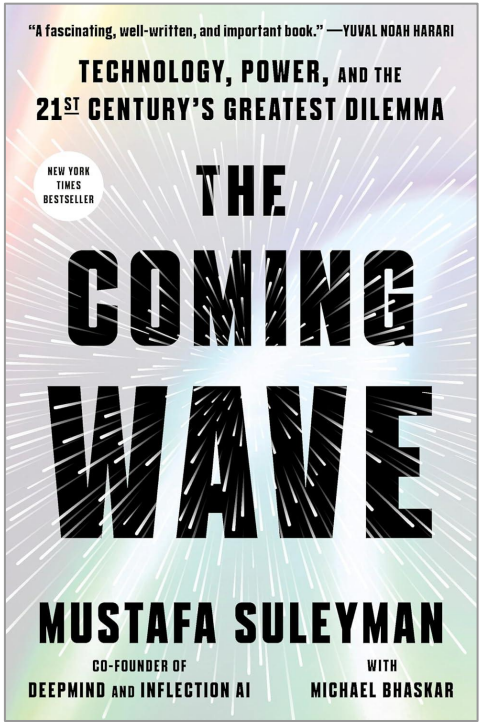


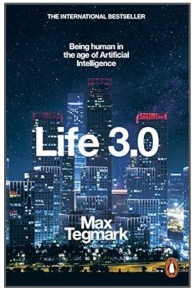
It is not the strongest species that survive,
nor the most intelligent,
but the ones ***most responsive to change***

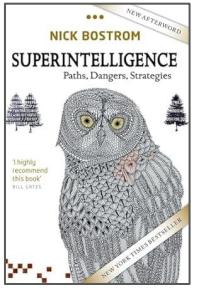
Charles Darwin, On the Origin of Species

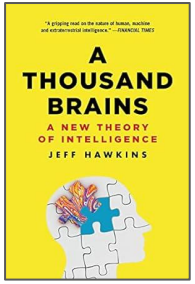
Book recomendations

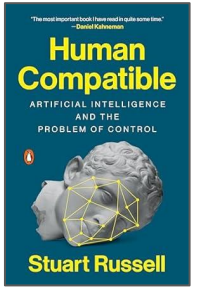












Q&A