UNIVERSITÀ DI PAVIA

Marco Piastra

Bias e Debiasing in Artificial Intelligence

ETICA PER LA PROGETTAZIONE
Almo Collegio Borromeo

13-04-2023

Prologue:
when the algorithm 'judges'

The COMPAS Algorithm

- **Correctional Offender Management Profiling for Alternative Sanctions**

  A proprietary algorithm by Northpointe Inc.
  (*now* Equivant Inc., https://www.equivant.com/)

It is a risk evaluation and decision support algorithm

It generates a risk score based on the answers to a questionnaire of 137 behavioral and psychological constructs (including criminal history)

The algorithm is patented and undisclosed

So is the dataset used for testing it

In July 2016, the Wisconsin Supreme Court ruled that COMPAS risk scores can be considered by judges during sentencing, but there must be warnings given to the scores to represent the tool's "limitations and cautions."

[images from https://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE.html]

Risk Assessment

PERSON

Name: [redacted] | Offender #: | DOB: [redacted]

[redacted] | Gender: Male | Marital Status: Single | Agency: DAI

ASSESSMENT INFORMATION

Case Identifier: [redacted] | Scale Set: Wisconsin Core - Community Language | Screener: [redacted] | Screening Date: [redacted]

Current Charges

☐ Homicide ☑ Weapons ☑ Assault ☐ Arson
☐ Robbery ☐ Burglary ☐ Property/Larceny ☐ Fraud
☐ Drug Trafficking/Sales ☐ Drug Possession/Use ☐ DUI/OUIL ☑ Other
☐ Sex Offense with Force ☐ Sex Offense w/o Force

1. Do any current offenses involve family violence?
   ☑ No ☐ Yes
2. Which offense category represents the most serious current offense?
   ☐ Misdemeanor ☐ Non-violent Felony ☑ Violent Felony
3. Was this person on probation or parole at the time of the current offense?
   ☑ Probation ☐ Parole ☐ Both ☐ Neither
4. Based on the screener's observations, is this person a suspected or admitted gang member?
   ☐ No ☑ Yes
5. Number of pending charges or holds?
   ☑ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4+
6. Is the current top charge felony property or fraud?
   ☑ No ☐ Yes

Criminal History

Exclude the current case for these questions.

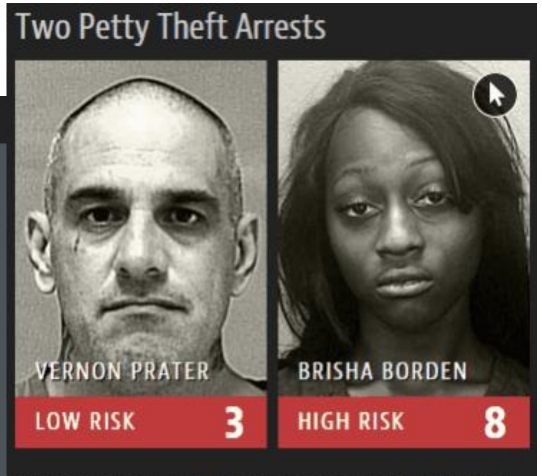
7. How many times has this person been arrested before as an adult or juvenile (criminal arrests only)?
   5
8. How many prior juvenile felony offense arrests?
   ☐ 0 ☐ 1 ☐ 2 ☐ 3 ☑ 4 ☐ 5+
9. How many prior juvenile violent felony offense arrests?
   ☐ 0 ☐ 1 ☑ 2+
10. How many prior commitments to a juvenile institution?
    ☐ 0 ☑ 1 ☐ 2+

[redacted] 8.1.18.12 [redacted]

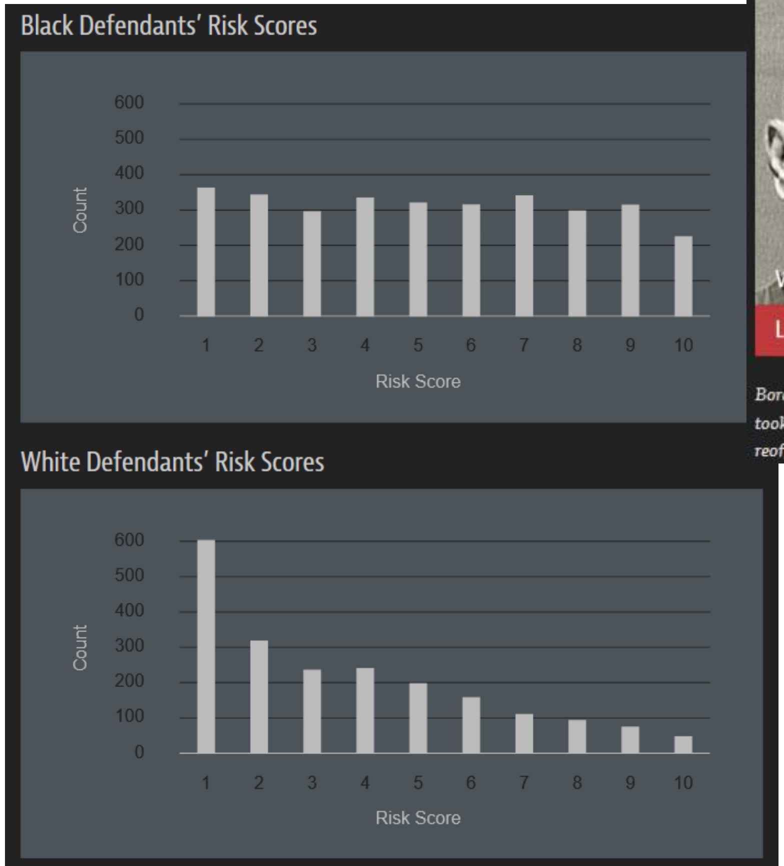
The COMPAS Algorithm

- **ProPublica claimed that the algorithm was biased** [2016]
  - *ProPublica is a nonprofit organization based in New York City*
  - *It is a newsroom that aims to produce investigative journalism in the public interest*

PROPUBLICA



Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.



These charts show that scores for white defendants were skewed toward lower-risk categories. Scores for black defendants were not. *(Source: ProPublica analysis of data from Broward County, Fla.)*

Their results, together with the methods and the dataset they used, were made public

- Such results have been criticized by Northpointe Inc. and other experts

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

The COMPAS Algorithm

- **No better than human judgement**

  A scientific study [2018] shows that COMPAS is not less reliable than a group of volunteers chosen at random on internet

ScienceAdvances Current Issue First release papers Archive About More

HOME > SCIENCE ADVANCES > VOL. 4, NO. 1 > THE ACCURACY, FAIRNESS, AND LIMITS OF PREDICTING RECIDIVISM

RESEARCH ARTICLE | RESEARCH METHODS

The accuracy, fairness, and limits of predicting recidivism

JULIA DRESSEL AND HANY FARID

SCIENCE ADVANCES • 17 Jan 2018 • Vol 4, Issue 1 • DOI: 10.1126/sciadv.aao5580

- **What is *fairness*, after all?**

  An article on The Washington Post [2016] puts into discussion the very notion of *fairness* in terms of mathematical objectivity

Sections The Washington Post Democracy Dies in Darkness €0.99 every four weeks Sign in

Monkey Cage

A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear.

By Sam Corbett-Davies, Emma Pierson, Avi Feller and Sharad Goel

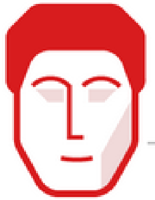
October 17, 2016

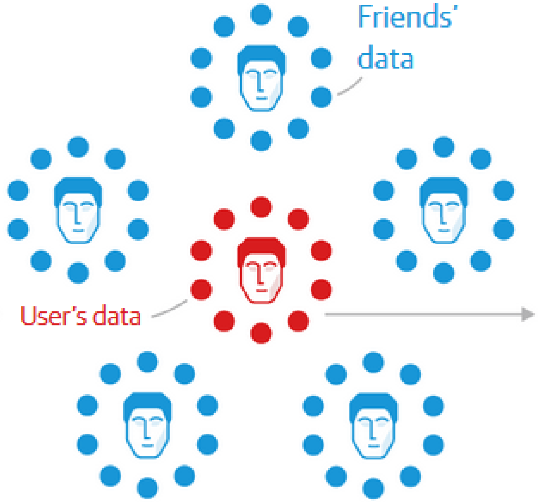
An aside:
There might be more patterns about us than we may want to admit…

Cambridge Analytica: how 50m Facebook records were hijacked

1
Approx. 320,000 US voters ('seeders') were **paid \$2-5 to take a detailed personality/ political test** that required them to log in with their Facebook account

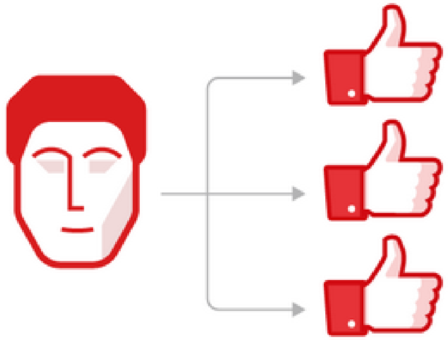
2
The app also **collected data such as likes and personal information** from the test-taker's Facebook account …



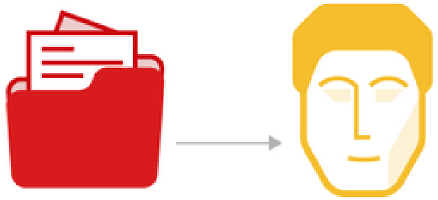


… as well their **friends'** data, amounting to over 50m people's raw Facebook data

3
The **personality quiz results** were paired with their Facebook data - such as **likes** - to seek out psychological patterns



4
Algorithms combined the data with other sources such as voter records to **create a superior set of records (initially 2m people in 11 key states*)**, with hundreds of data points per person



These individuals could then be targeted with **highly personalised advertising** based on their personality data

[Graphics from https://www.theguardian.com/technology/2018/mar/17/facebook-cambridge-analytica-kogan-data-algorithm]

The Cambridge Analytical Scandal

- **Scientific foundations: the method**

  Two well-known articles by Kosinski et al.

PNAS

Private traits and attributes are predictable from digital records of human behavior

Michal Kosinski[a,1], **David Stillwell**[a], **and Thore Graepel**[b]

[a]Free School Lane, The Psychometrics Centre, University of Cambridge, Cambridge CB2 3RQ United Kingdom; and [b]Microsoft Research, Cambridge CB1 2FB, United Kingdom

Edited by Kenneth Wachter, University of California, Berkeley, CA, and approved February 12, 2013 (received for review October 29, 2012)

PNAS

Computer-based personality judgments are more accurate than those made by humans

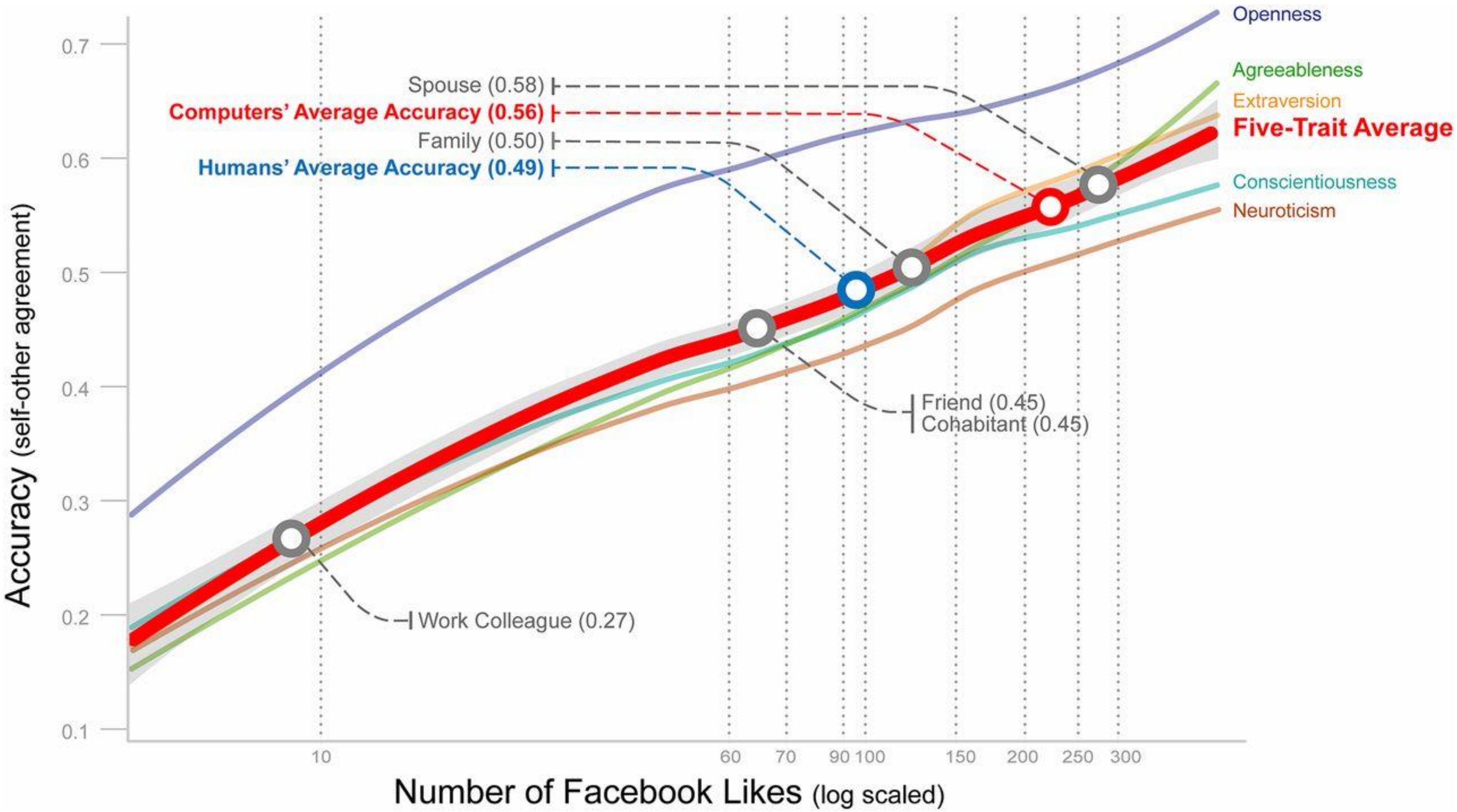
Wu Youyou[a,1,2], **Michal Kosinski**[b,1], **and David Stillwell**[a]

[a]Department of Psychology, University of Cambridge, Cambridge CB2 3EB, United Kingdom; and [b]Department of Computer Science, Stanford University, Stanford, CA 94305

Edited by David Funder, University of California, Riverside, CA, and accepted by the Editorial Board December 2, 2014 (received for review September 28, 2014)

The Cambridge Analytical Scandal

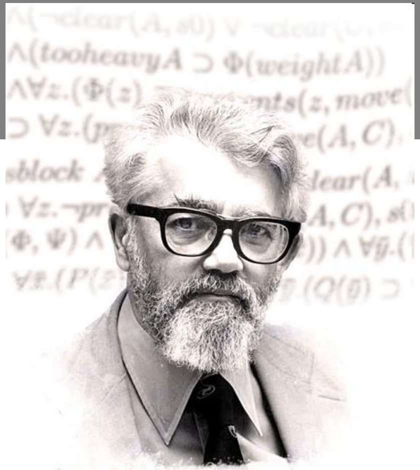
- **The “Big Five” personality traits are predictable from Facebook likes**



Wu Youyou et al. PNAS 2015;112:4:1036-1040

At the edge of algorithms: Artificial Intelligence

"Artificial Intelligence" (first appearance of the term)



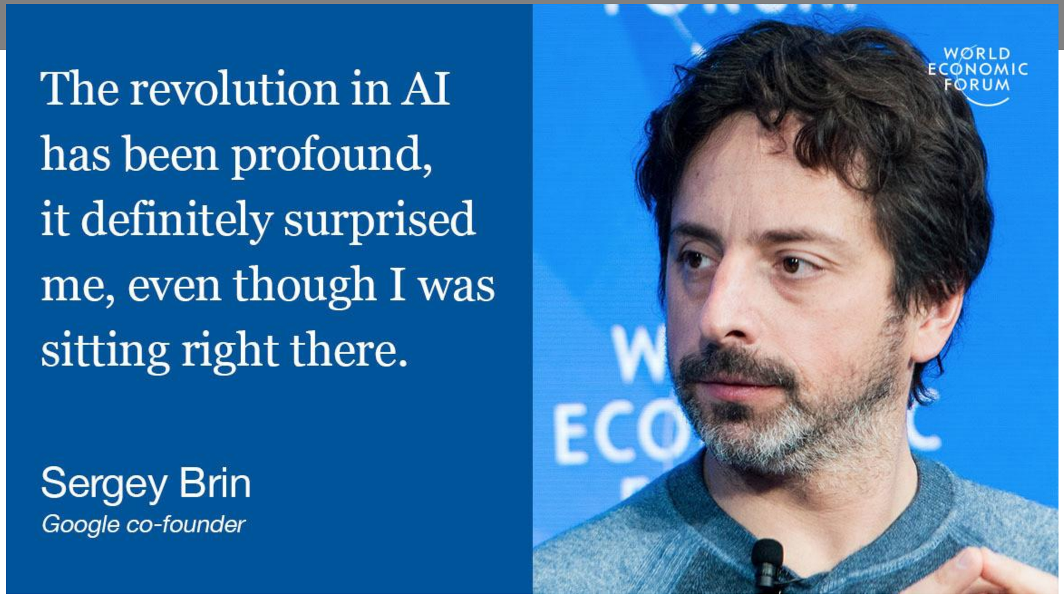
[Image from Wikipedia]

"We propose that a two-month, ten man study
of **artificial intelligence** carried out during the summer of 1956 [...]

The study is to proceed on the basis of the conjecture
that every aspect of learning or any other feature of **intelligence**
can in principle be ***so precisely described***
that a machine can be made to ***simulate*** it. […]"

[John McCarthy et al., 1955, *emphasis added*]

How did it go, in reality?



- Sergey Brin [Google Co-Founder, January 2017]

  "I didn't pay attention to it [i.e. Artificial Intelligence] at all, to be perfectly honest."

  *"Having been trained as a computer scientist in the 90s, everybody knew that AI didn't work.
  People tried it, they tried neural nets and none of it worked."*

[Quote and image from https://www.weforum.org/agenda/2017/01/google-sergey-brin-i-didn-t-see-ai-coming/]

AI on the Rise: is that Good?

The New York Times

Artificial Intelligence › | An Unsettling Chat With Bing | Read the Conversation | How Chatbots Work | Spotting A.I.-Generated Text

Google Researcher Says She Was Fired Over Paper Highlighting Bias in A.I.

Timnit Gebru, one of the few Black women in her field, had voiced exasperation over the company's response to efforts to increase minority hiring.

Give this article | 276

Timnit Gebru, a respected researcher at Google, questioned biases built into artificial intelligence systems. Cody O'Loughlin for The New York Times

By Cade Metz and Daisuke Wakabayashi

Dec. 3, 2020

[Quote from https://www.nytimes.com/2020/12/03/technology/google-researcher-timnit-gebru.html1

Artificial Intelligence Hysteria?

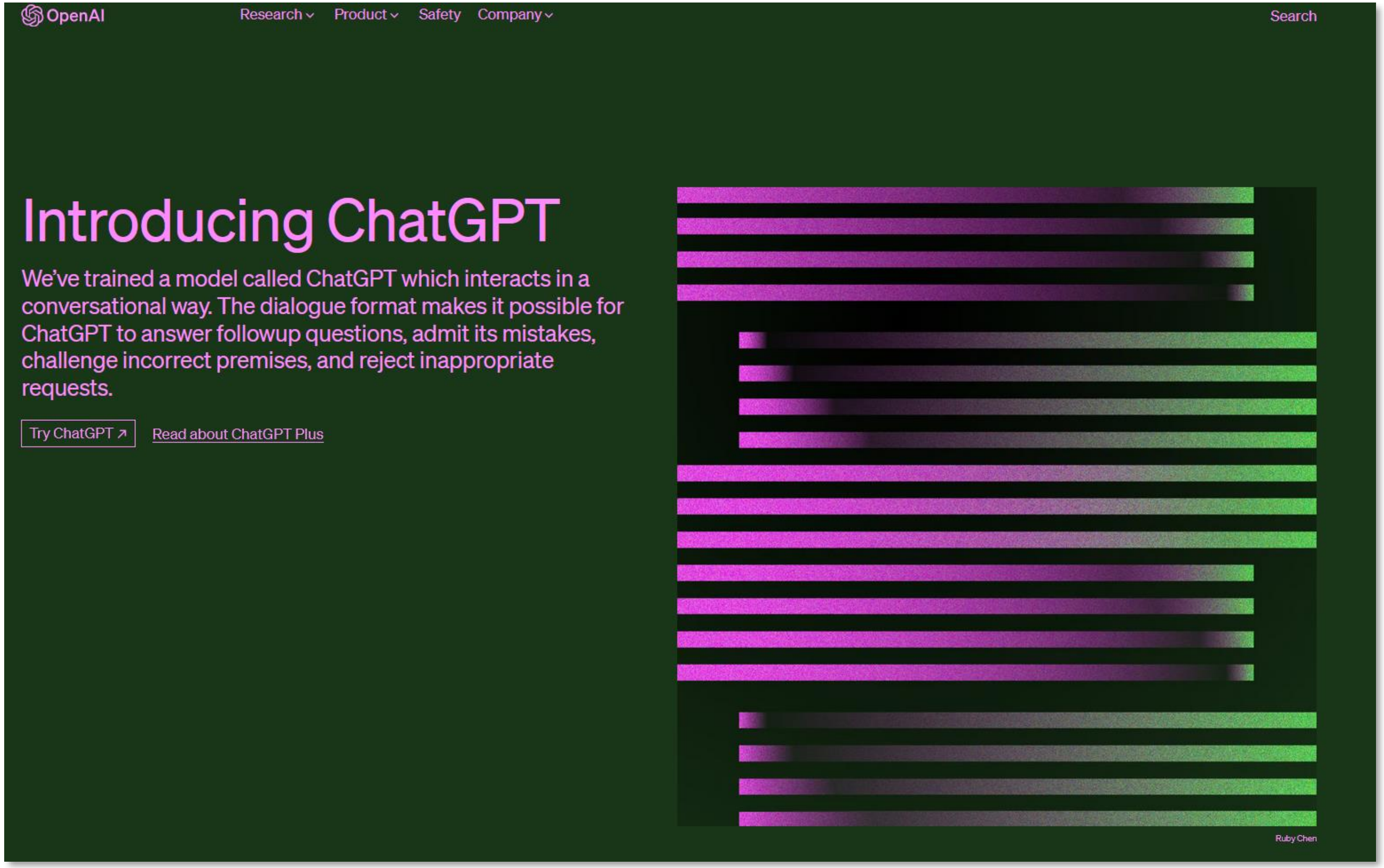
AI isn't as scary as we imagine. AndreyZH/Shutterstock

The reality of AI is currently very different, particularly when you look at the threat of automation. Back in 2013, researchers estimated that, in the following ten to 20 years, 47% of jobs in the US could be automated. Six years later, instead of a trend towards mass joblessness, we're in fact seeing US unemployment at a historic low.

Current AI is good at finding patterns in large datasets, and not much else.

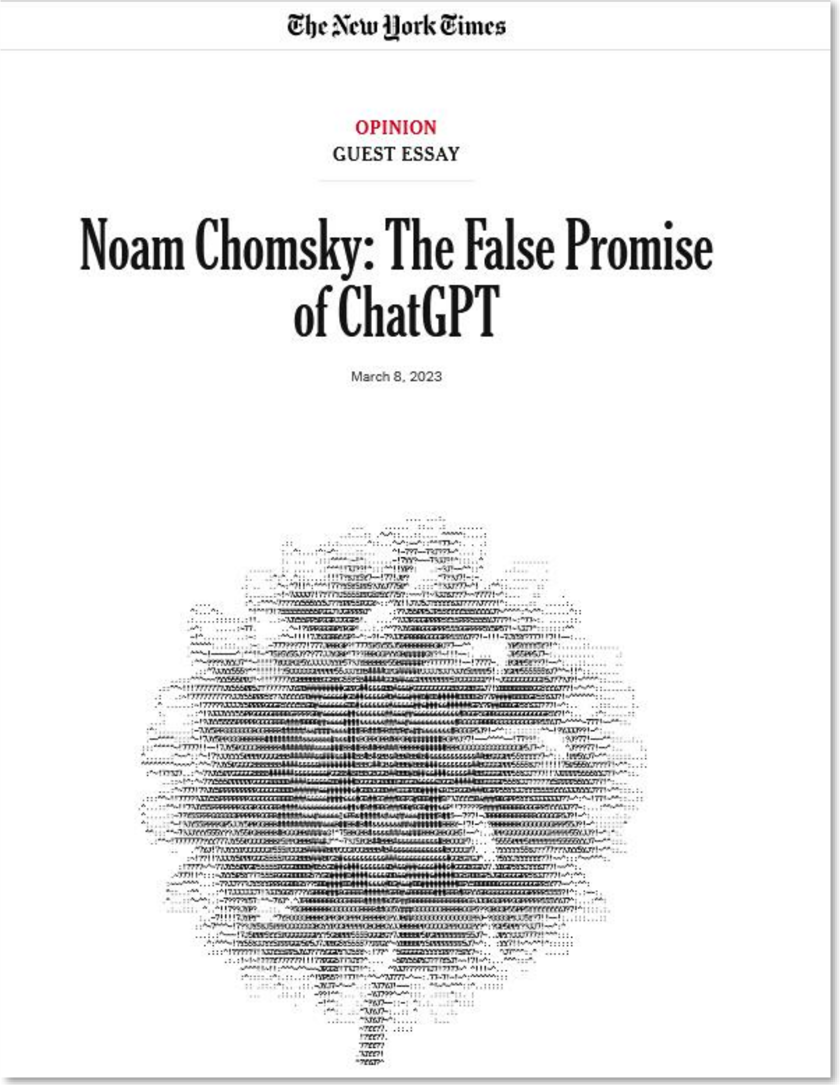
[Quote from https://theconversation.com/ais-current-hype-and-hysteria-could-set-the-technology-back-by-decades-120514]

Is Artificial Intelligence Here to Stay?



[Image from https://openai.com/blog/chatgpt, 09/03/2023]

Is Artificial Intelligence Intelligent?


The New York Times

OPINION
GUEST ESSAY

Noam Chomsky: The False Promise of ChatGPT

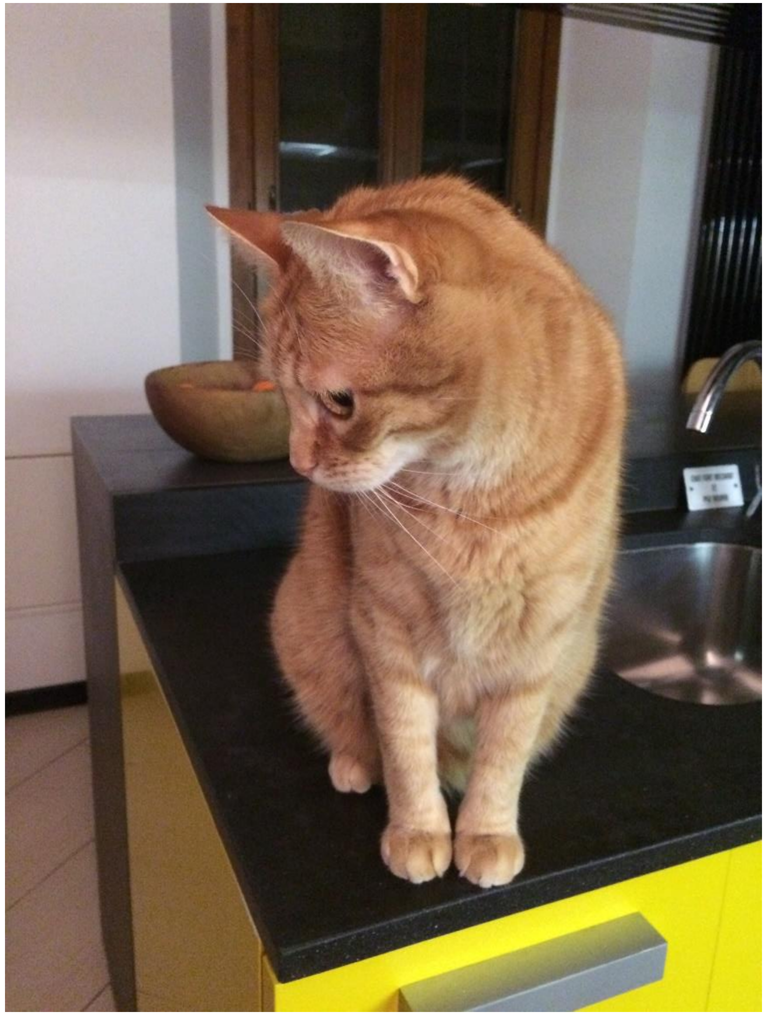
March 8, 2023

[Image from https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html]

Looking Inside:
A function to say 'Cat!'

Artificial Perception

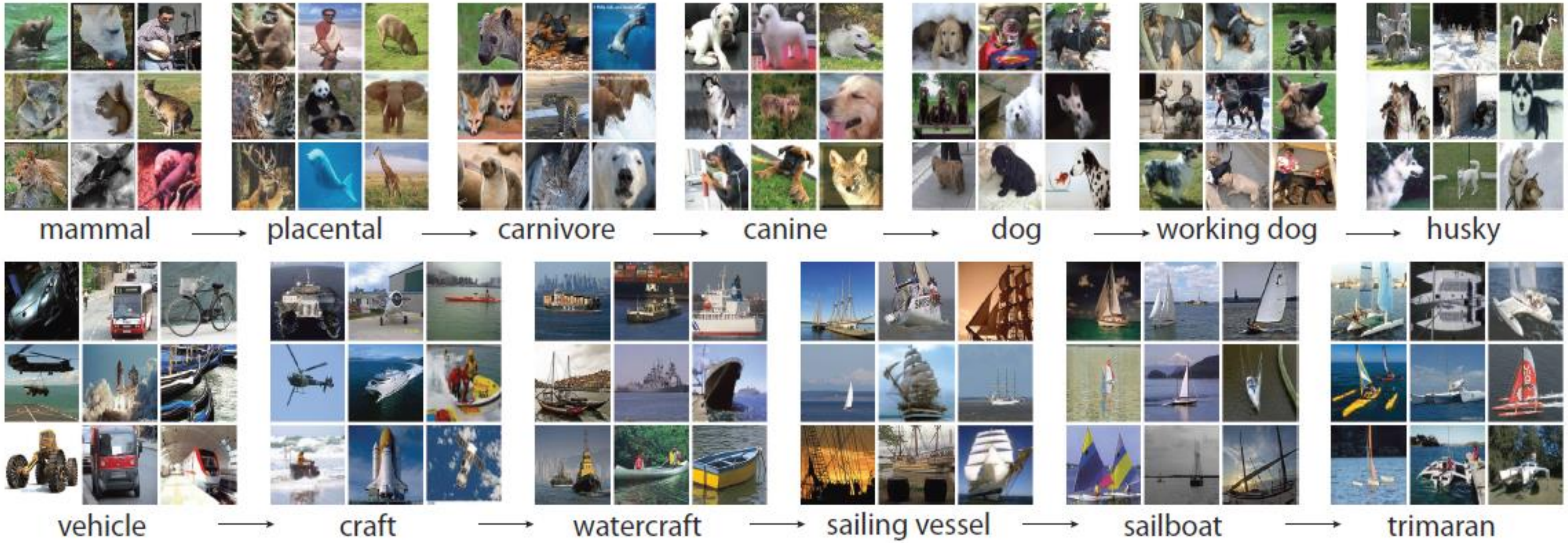
*Is there a cat
in this picture?*



[this is *my* cat, Rabarbaro]

ImageNet Challenge

- The ImageNet Large Scale Visual Recognition Challenge



1,461,406 full resolution images

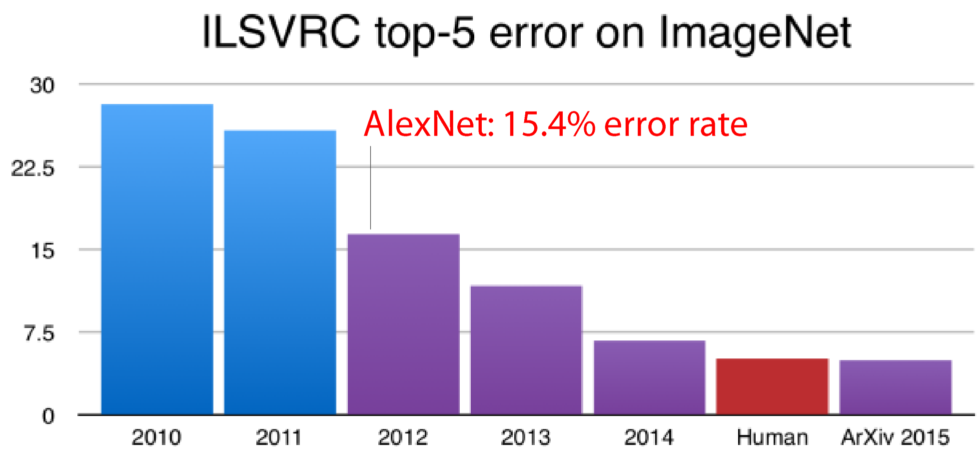
Complex and multiple textual annotation,
hierarchy of 1000 object classes along several dimensions

The image classification challenge was run annually from 2010 to 2017

[figures from www.nvidia.com]

ImageNet Challenge

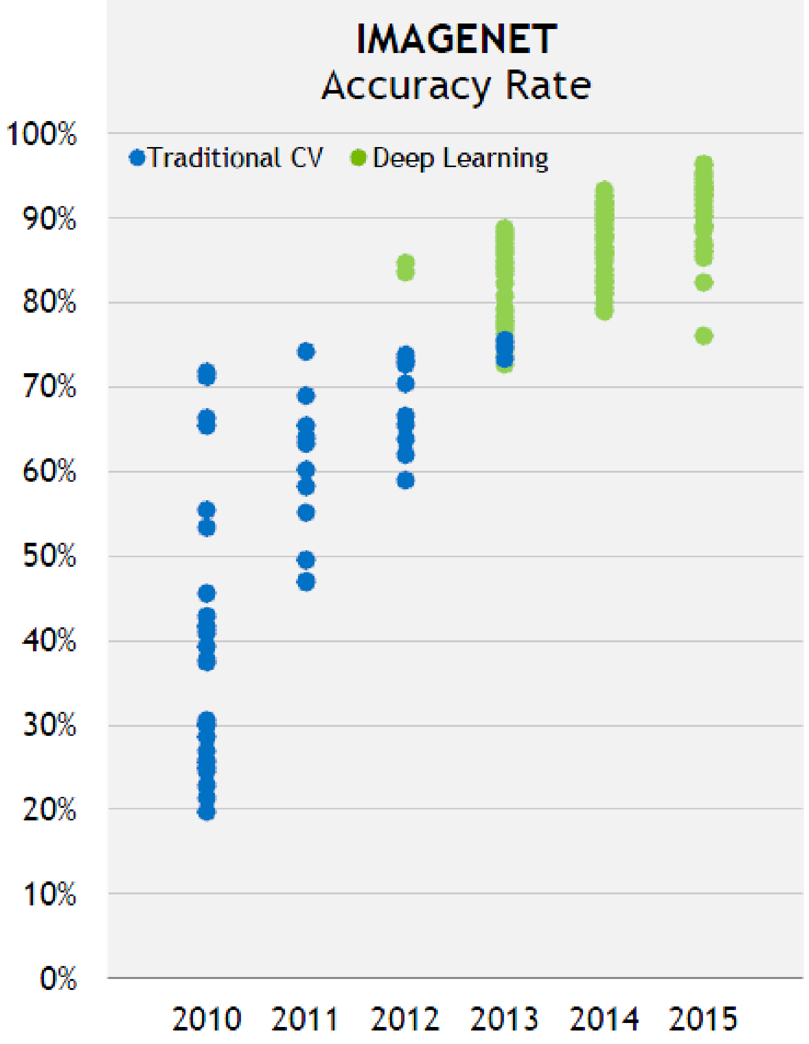
- The ImageNet Large Scale Visual Recognition Challenge



1,461,406 full resolution images

Complex and multiple textual annotation,
hierarchy of 1000 object classes along several dimensions

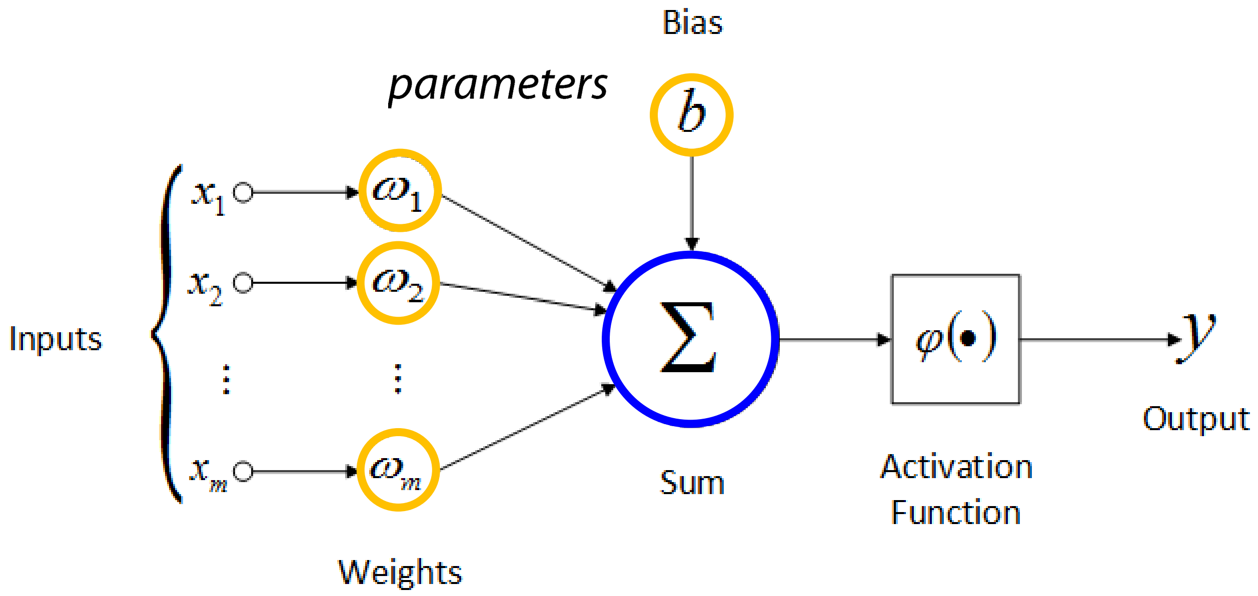
The image classification challenge was run annually from 2010 to 2017

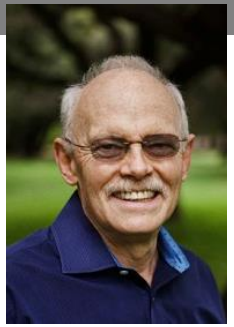


[figures from www.nvidia.com]

How They Did It:
Deep Neural Networks

Artificial Neural Networks





[Images from Wikipedia]

[Rumelhart, D.E., J.L. McClelland 1986]

- ***Basic assumption***

  Mental phenomena can be described
  by interconnected networks of simple and often uniform units

Artificial Neural Networks

- **From *shallow* to *deep* networks**

A feed-forward neural network with one hidden layer

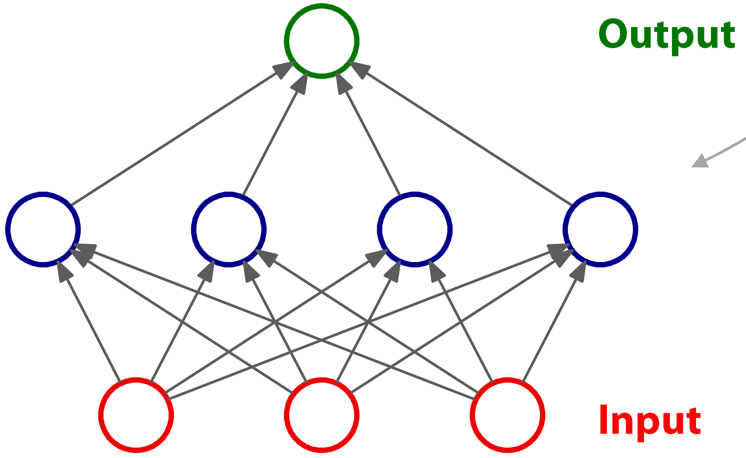
$$\tilde{y} = \boldsymbol{w} \cdot g(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}) + b$$

Deep Learning systems (e.g. TensorFlow, PyTorch) use this representation

It can approximate any target function

$$y = f^*(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathbb{R}^d$$

(given enough units and proper parameters)



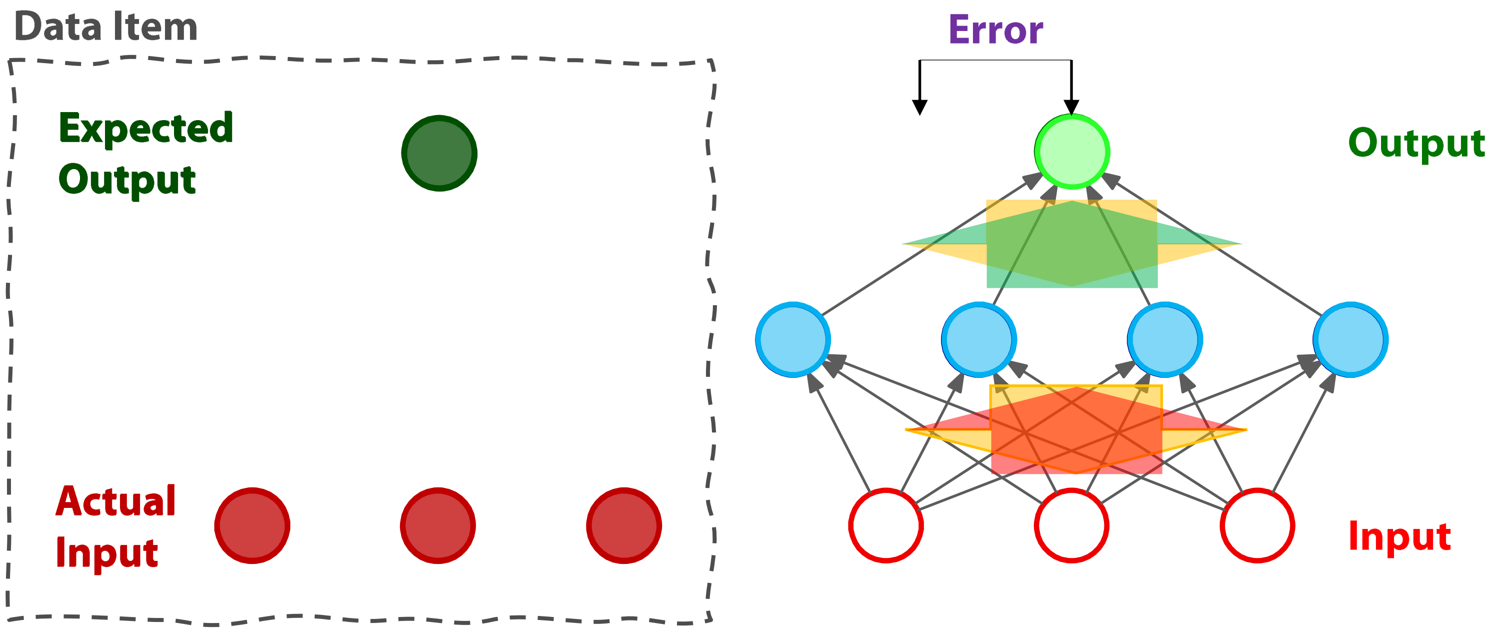
The two representations are equivalent

- **Learning is a *parameter optimization* process**

Using a large dataset of input-output pairs (*data items*)

$$\tilde{y} = \boldsymbol{w} \cdot g(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}) + b$$

Feed Data Item(s)
Improve
Repeat
Several million times ...



Propagate Input to compute Output

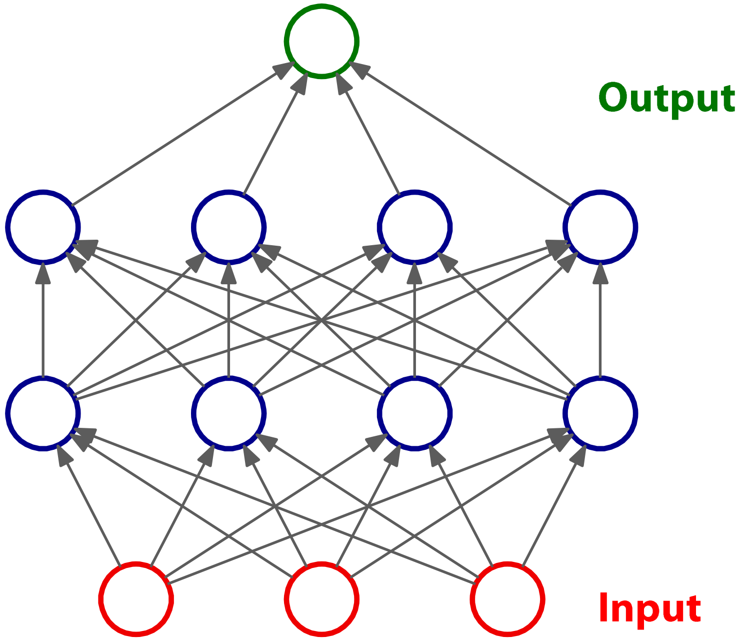
Propagate Error to improve *parameters*

Deep Neural Networks

- **From *shallow* to *deep* networks**

  A feed-forward neural network with *two* hidden layers

$$\tilde{y} = \boldsymbol{w} \cdot g(\boldsymbol{W}^{[2]} g(\boldsymbol{W}^{[1]} \boldsymbol{x} + \boldsymbol{b}^{[1]}) + \boldsymbol{b}^{[2]}) + b$$

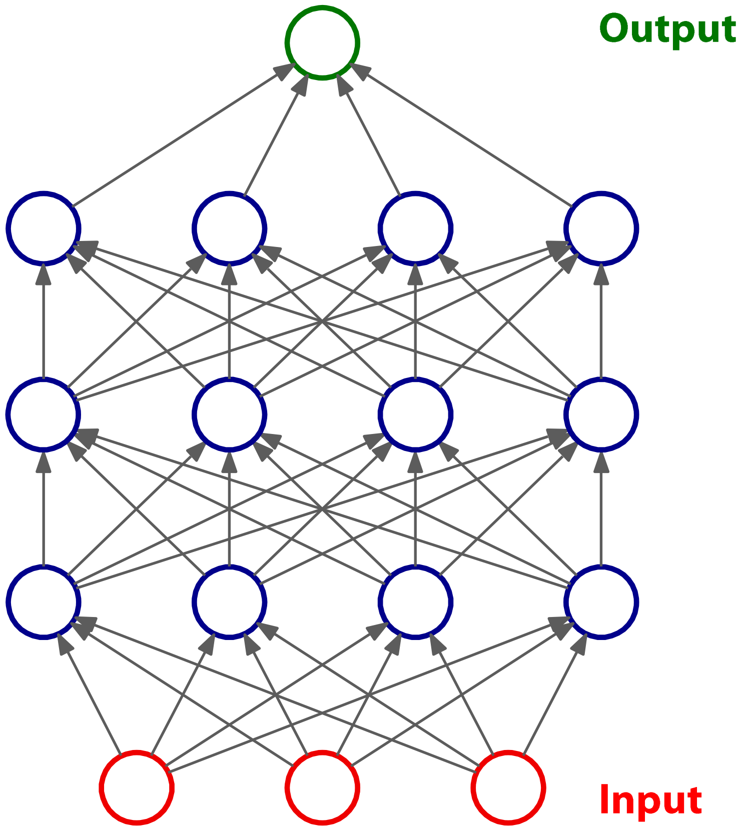


Deep Neural Networks

- **From *shallow* to *deep* networks**

  A feed-forward neural network with <u>*three*</u> hidden layers

$$\tilde{y} = \boldsymbol{w} \cdot g(\boldsymbol{W}^{[3]} g(\boldsymbol{W}^{[2]} g(\boldsymbol{W}^{[1]} \boldsymbol{x} + \boldsymbol{b}^{[1]}) + \boldsymbol{b}^{[2]}) + \boldsymbol{b}^{[3]}) + b$$



Deep Convolutional Neural Networks (DCNN)

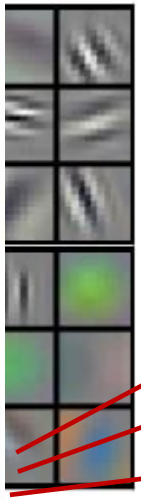
- ***AlexNet*** [Krizhevsky, Sutskever & Hinton, 2012]

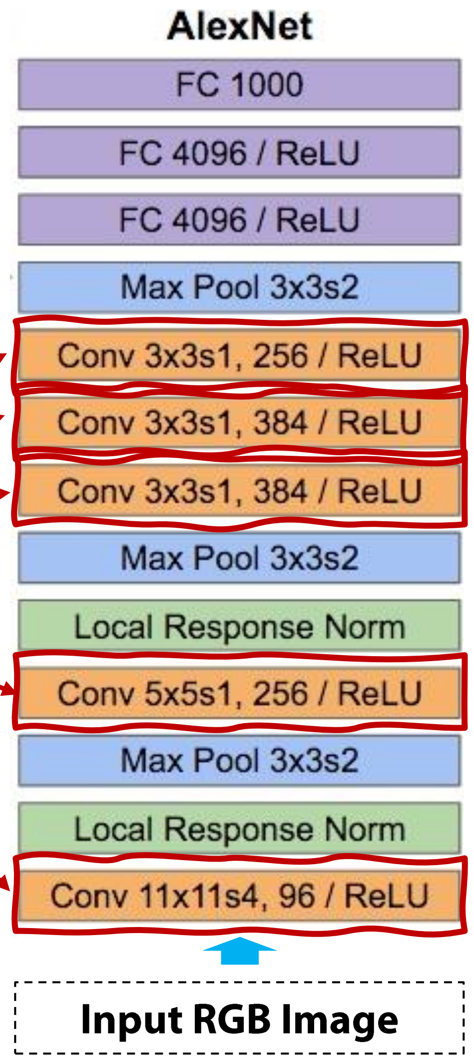
Layer 4: 384 deep convolutive filters

Layer 1: 9

Layer 3: 384 deep convo

Layer 2: 256 deep convo





[images from http://cs231n.github.io/convolutional-networks/]

Object (and People) Real-Time Detection

- *Deep Convolutional Neural Networks have evolved since then ...*

Now these system can identify objects and persons from videos, in real time

NOTE:
According to the recent
EU Proposal for a Regulation
*about AI, **remote biometric identification** (RBI)*
in public places will require
a special authorization

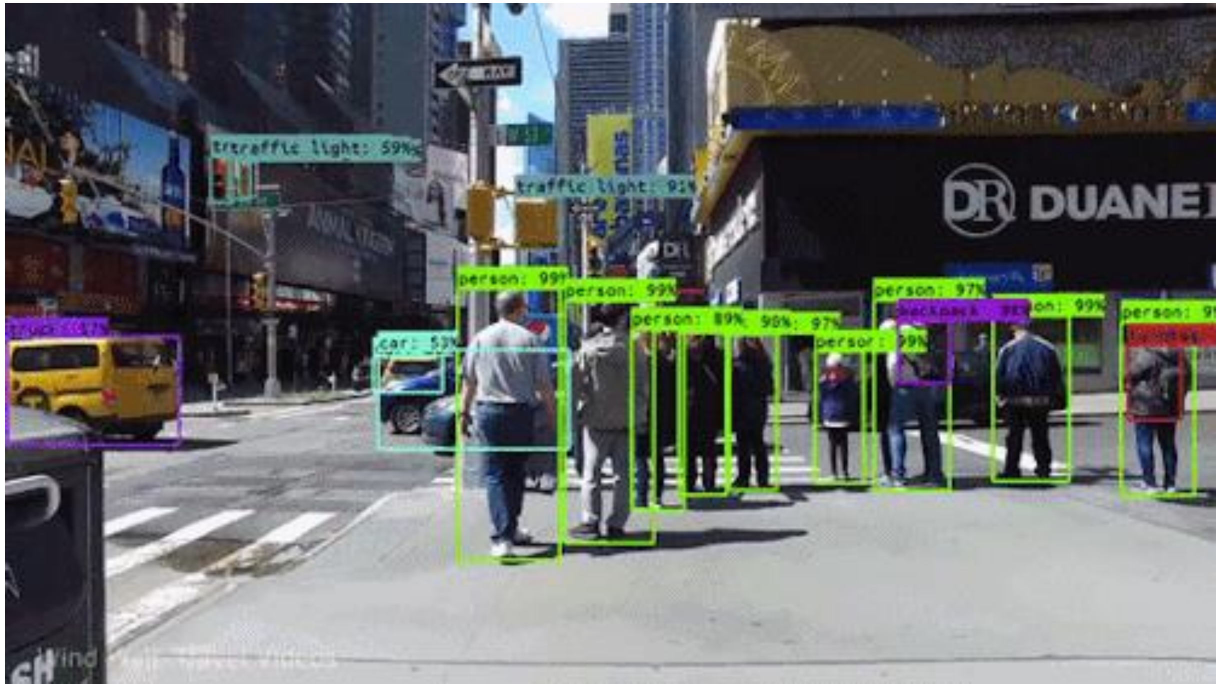


Image from: https://sgu.ac.id/id/computer-vision-artificial-intelligence-why-is-it-important/

Image Segmentation

- *Deep Convolutional Neural Networks have evolved since then ...*

They can perform a complete scene analysis, from videos, in real time

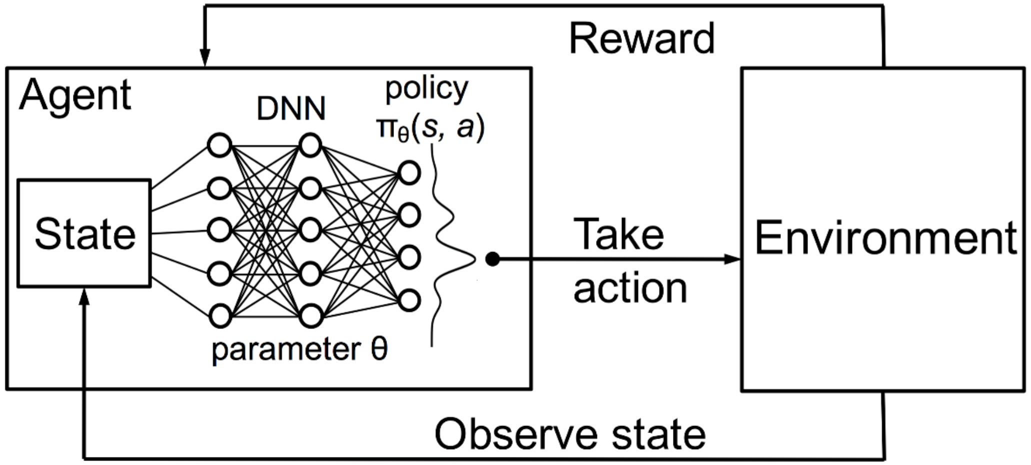
At present, DCNN work on a frame-by-frame basis

Image from: https://sgu.ac.id/id/computer-vision-artificial-intelligence-why-is-it-important/

Well, it's just a <u>function</u> *anyway …*

Deep Reinforcement Learning (DRL)

- ***A Deep Neural Network learns a <u>policy</u>***



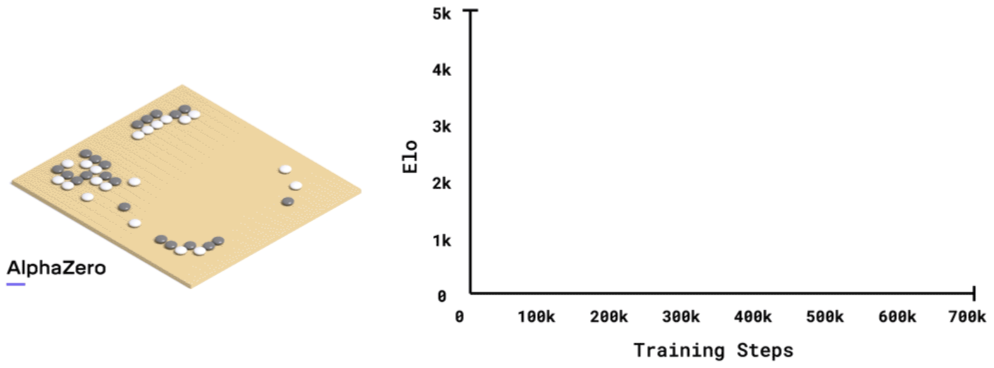
The agent interacts with an environment (it could be a copy of itself)

*It selects an **action** in each **state** and receives a **reward** (possibly deferred) as a function of the results obtained*

The DRL system optimizes its <u>policy</u>

Image from: https://deepmind.com/blog/article/alphazero-shedding-new-light-grand-games-chess-shogi-and-go

Some famous game-playing algorithms
are heavily reliant
on the experience of human players



- **AlphaZero learns to play *by itself***
  [2018, D. Silver, et al. (13 authors), https://science.sciencemag.org/content/362/6419/1140.full]

Basic Knowledge Only

It just knows the basic rules of the games

Learning via Self-Play

It plays against a (frozen) copy of itself

MCTS and DCNN in a closed loop

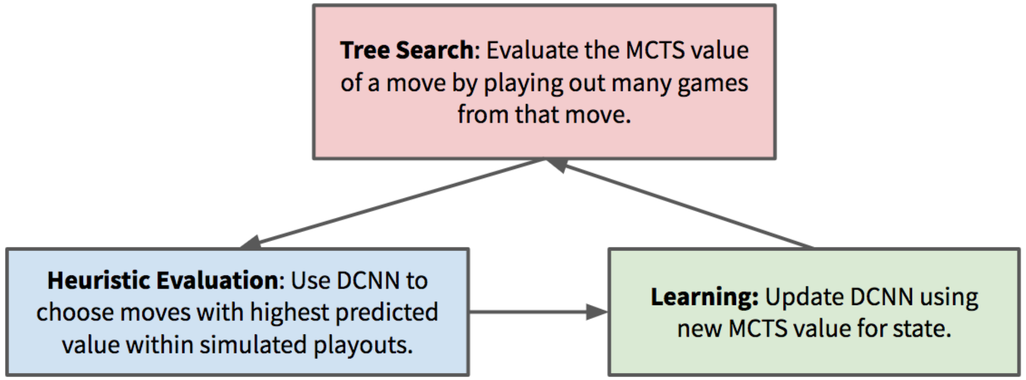
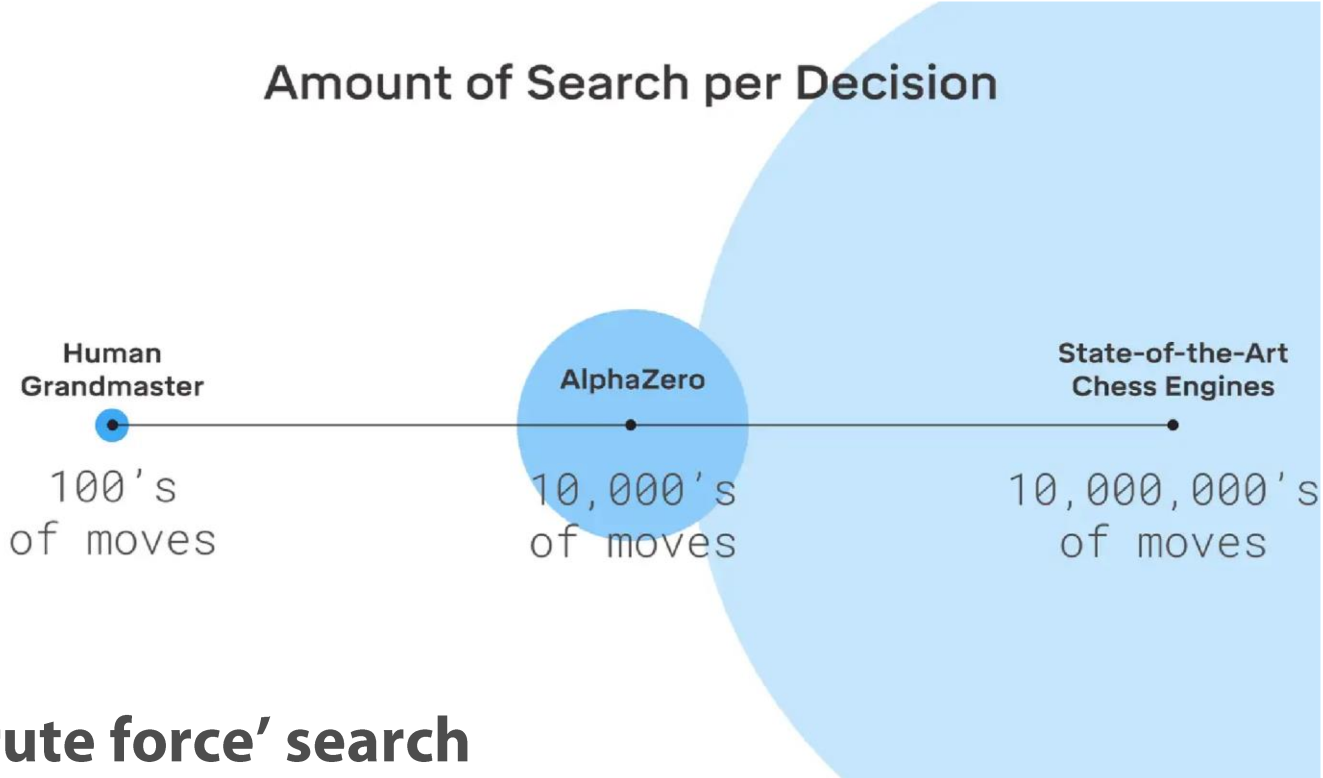


Image from: https://nikcheerla.github.io/deeplearningschool/2018/01/01/AlphaZero-Explained/

Beyond Emulating Humans: AlphaZero (2018)

Image from: https://deepmind.com/blog/article/alphazero-shedding-new-light-grand-games-chess-shogi-and-go



- **AlphaZero uses much less 'brute force' search**
    - When playing, the search process is driven by its neural network
        - *It acts like a memory of past experiences*
    - While training, it learns through a huge amount of self-playing

GPT-3 (2020)

Image from https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3

Support The Guardian

Available for everyone, funded by readers

Contribute → Subscribe →

My account

The Guardian

Opinion Artificial intelligence (AI)

A robot wrote this entire article. Are you scared yet, human?

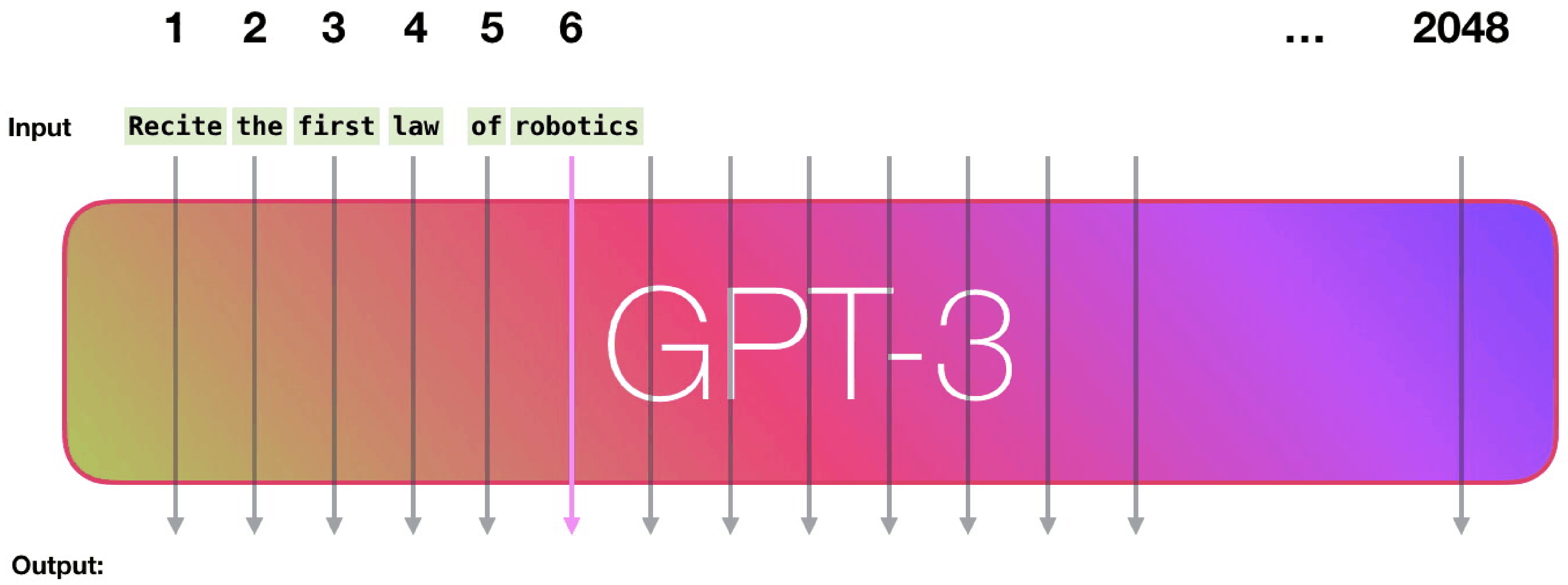
GPT-3

We asked GPT-3, OpenAI's powerful new language generator, to write an essay for us from scratch. The assignment? To convince us robots come in peace

- For more about GPT-3 and how this essay was written and edited, please read our editor's note below

Tue 8 Sep 2020 09.45 BST

70,298 1,188



One of the biggest Neural Networks yet

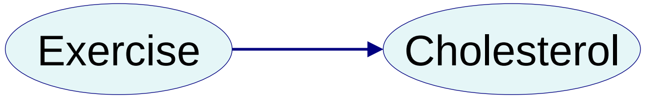
GPT-3 has 175 Billion parameters
(AlexNet has 64 Million)

Image from http://jalammar.github.io/how-gpt3-works-visualizations-animations/

Fairness and Probability

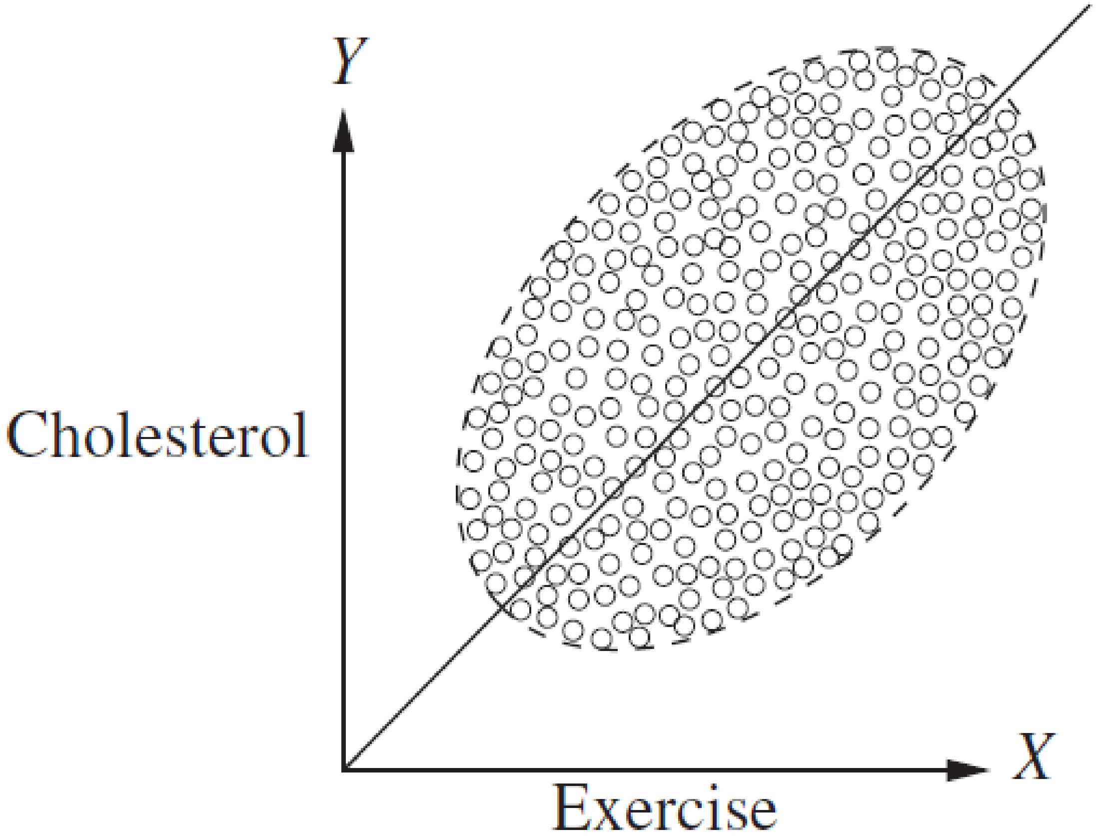
Causes and Effects: the Simpson's Paradox [1922]

- Does physical exercise prevent cholesterol?



Apparently not: correlation is *positive*

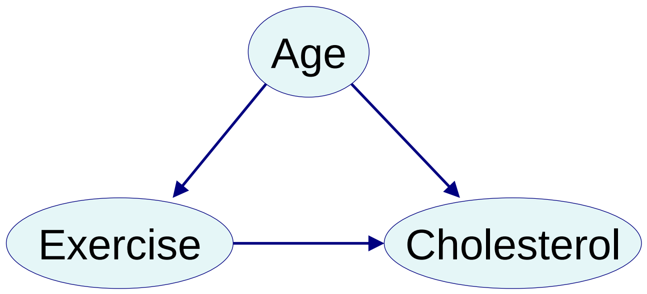
In words:
<u>more</u> physical exercise corresponds to *(causes?)*
<u>more</u> cholesterol ...



[Image from Pearl, J. et al., "Causal Inference in Statistics: A Primer", Wiley, 2016]

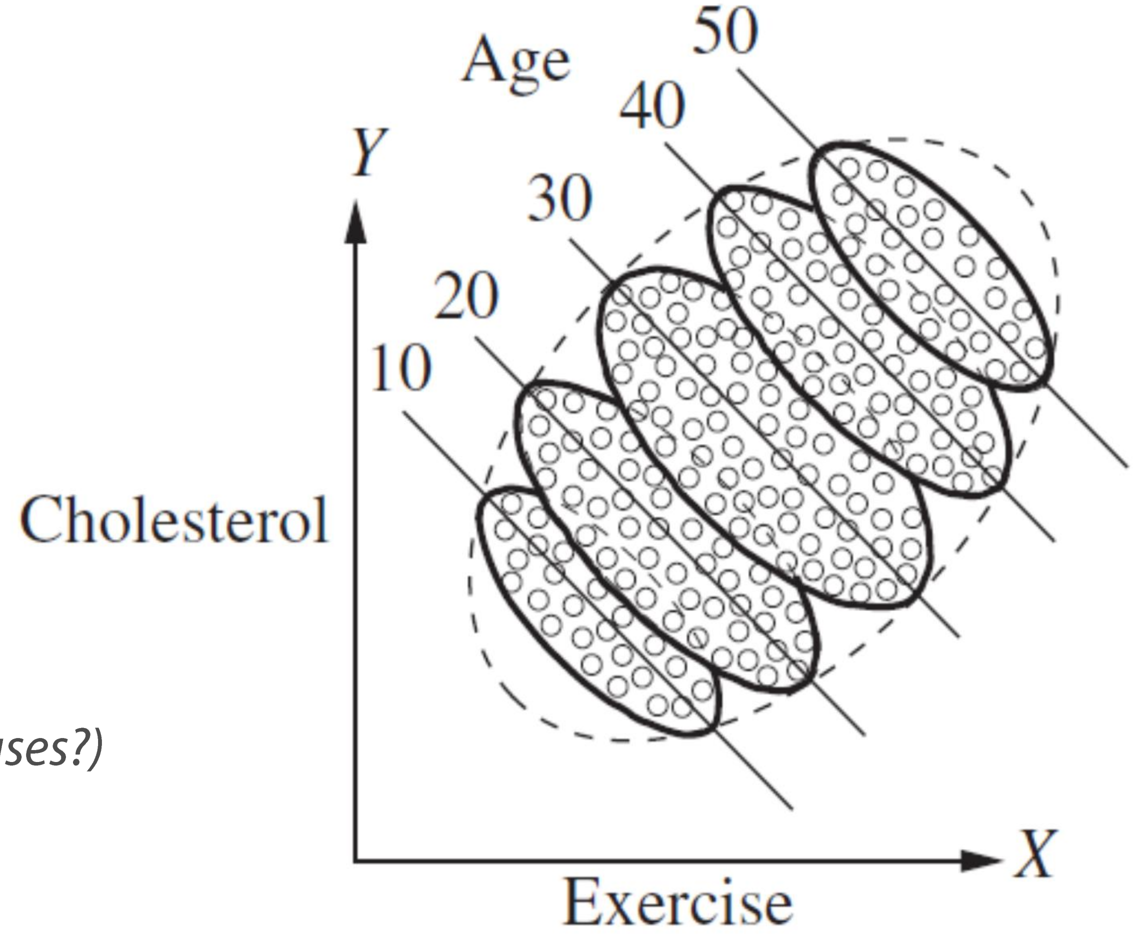
- Does physical exercise prevent cholesterol?

Maybe, if we consider another variable...



Correlation in each Age subgroups is *negative*

In words:
in each age group, more exercise corresponds to *(causes?)* less cholesterol ...

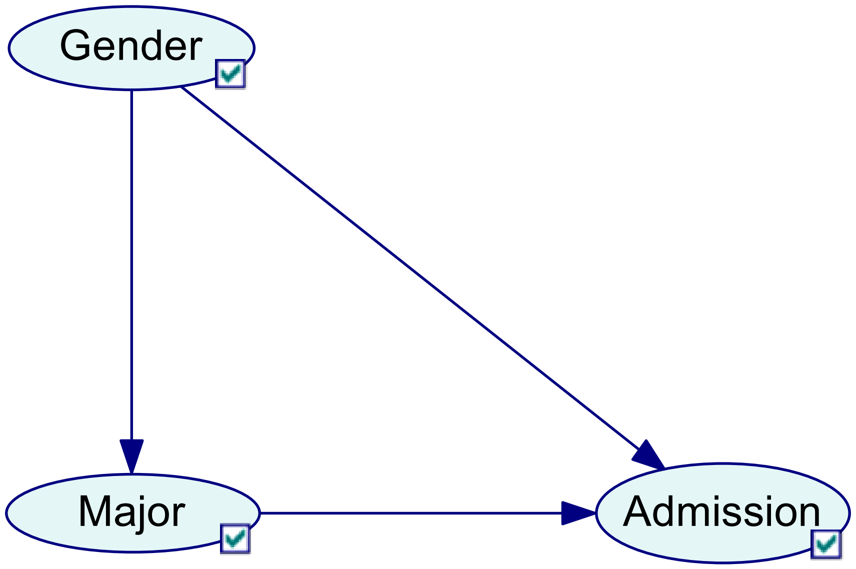


[Image from Pearl, J. et al., "Causal Inference in Statistics: A Primer", Wiley, 2016]

A Real-World Example: Student Admissions at UC Berkeley

A public domain dataset: all 12,763 applicants to UC-Berkeley's undergraduate programs in Fall 1973

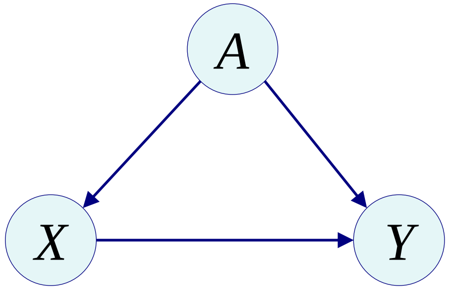
- ***Does Gender matter, for admission?***



Gender also plays a role in the selection of Major, the academic discipline for which the student applies

(see example in GeNIe)

- A generic *predictor*, i.e., an algorithm



A is a sensitive attribute

X is other attributes (there could be many)

Y is the actual outcome

$\tilde{Y}$ is the predicted outcome, given A and X

Demographic Parity

$< \tilde{Y} \perp A >$ "Prediction does not depend on the sensitive attribute"

Predictive Parity

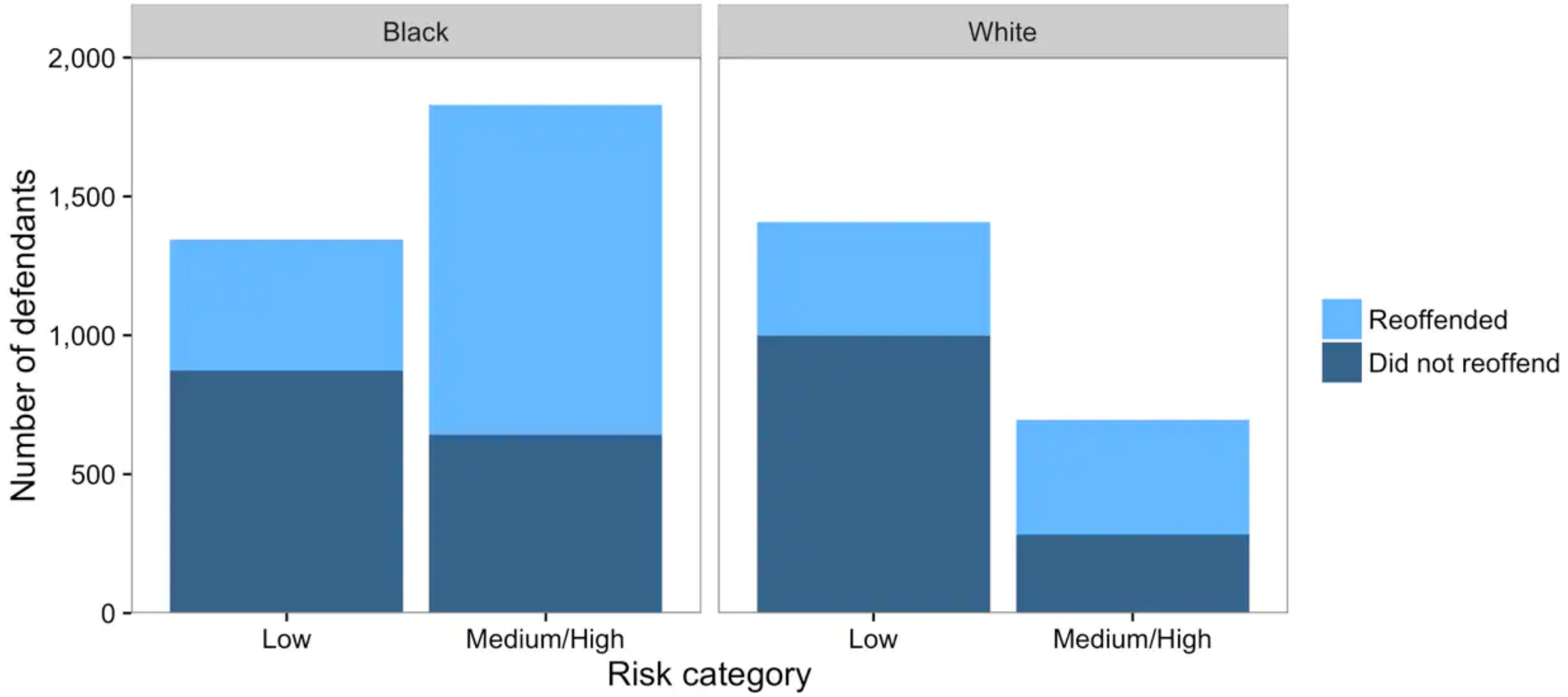
$< Y \perp A \mid \tilde{Y} >$ "Equal error rates, altogether"

Equal False Positive and False Negative Rates

$< \tilde{Y} \perp A \mid Y >$ "Prediction errors do not depend on the sensitive attribute"

The COMPAS Algorithm Revisited

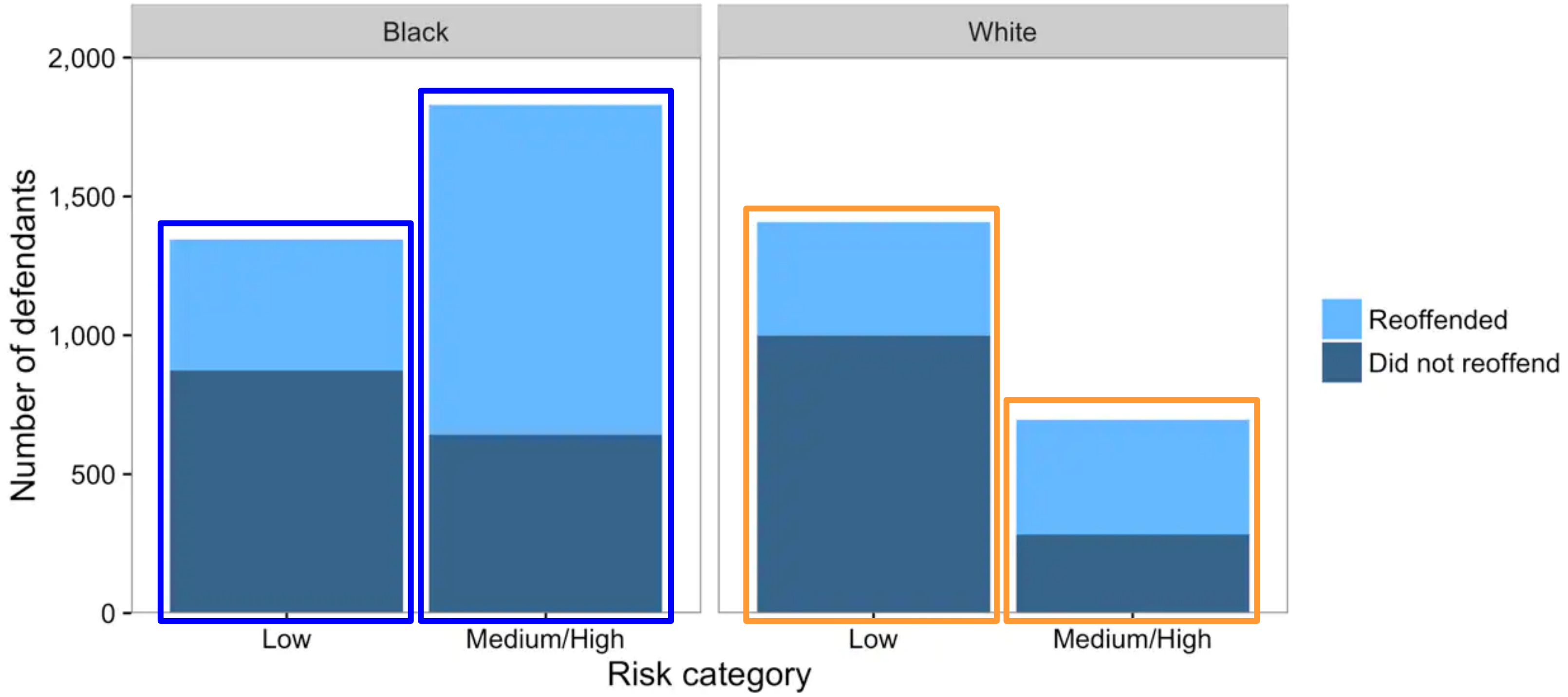
A is *race* (black vs. white)
X is other attributes (omitted)
Y is the actual outcome: did the subject *reoffend*?
$\tilde{Y}$ is the predicted outcome: which *risk category* has been assigned to the defendant?



[Image from https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/]

The COMPAS Algorithm Revisited

A is *race* (black vs. white)
X is other attributes (omitted)
Y is the actual outcome: did the subject *reoffend*?
$\tilde{Y}$ is the predicted outcome: which *risk category* has been assigned to the defendant?



Demographic Parity

$< \tilde{Y} \perp A >$

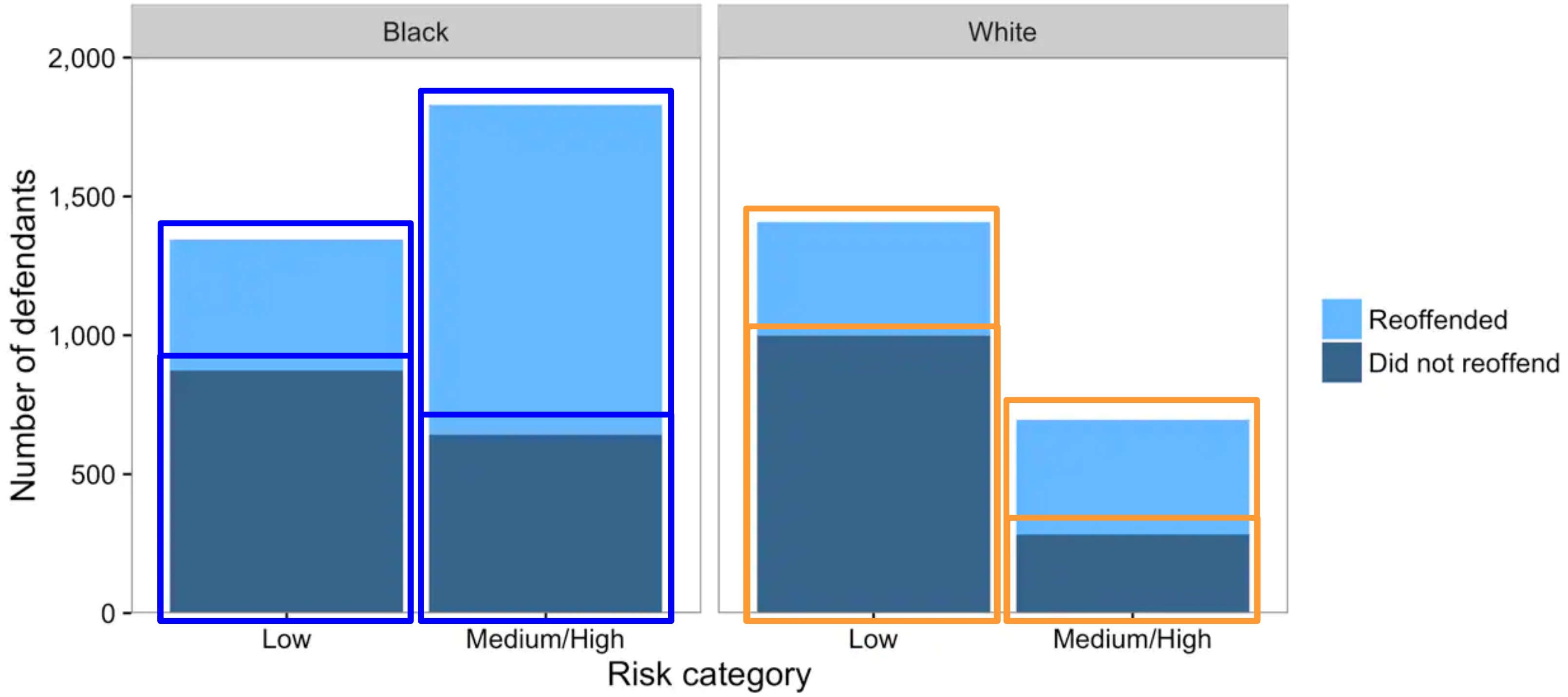
The height proportion of the two columns for each group should be the same

False

[Image from https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/]

The COMPAS Algorithm Revisited

A is *race* (black vs. white)
X is other attributes (omitted)
Y is the actual outcome: did the subject *reoffend*?
$\tilde{Y}$ is the predicted outcome: which *risk category* has been assigned to the defendant?



Predictive Parity

$< Y \perp A \mid \tilde{Y} >$

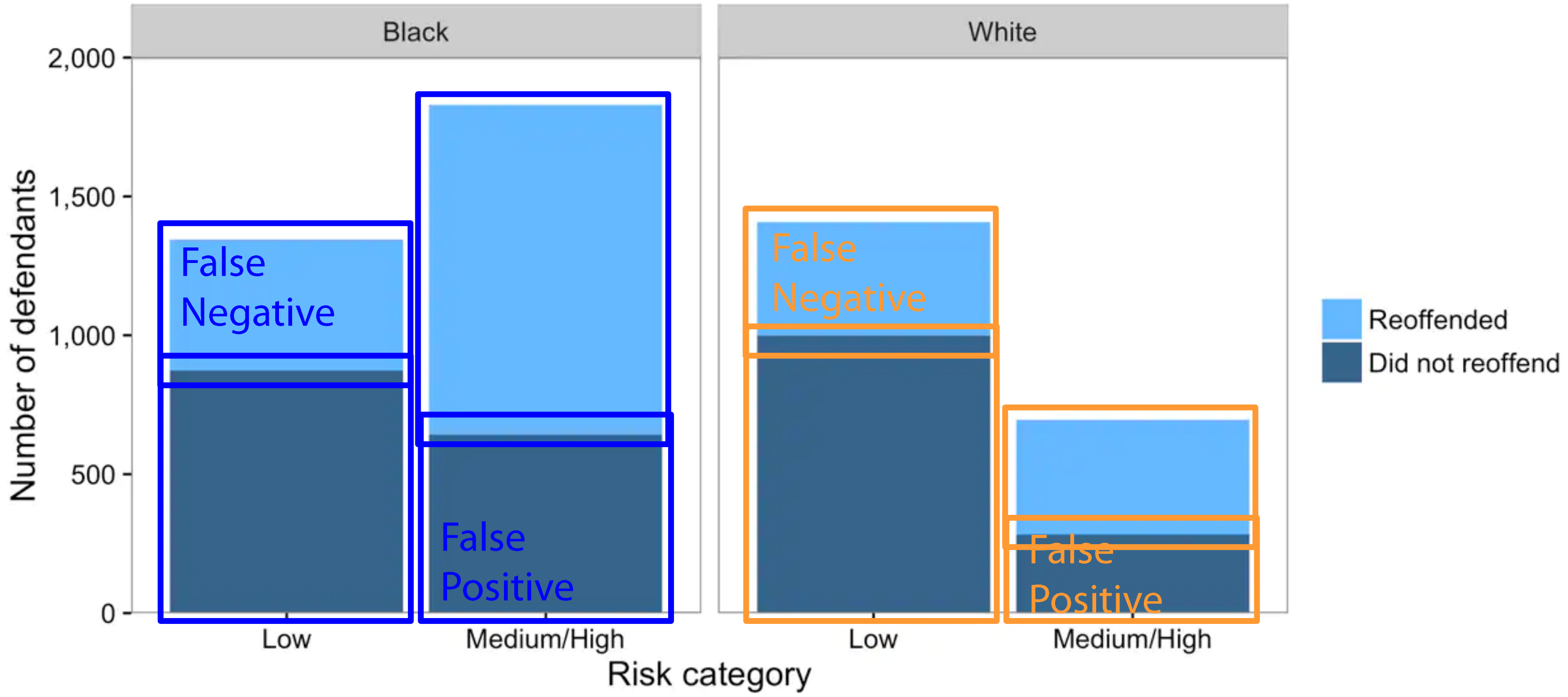
The prediction should be wrong equally often

***True** (more or less)*

[Image from https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/]

The COMPAS Algorithm Revisited

A is *race* (black vs. white)
X is other attributes (omitted)
Y is the actual outcome: did the subject *reoffend*?
$\tilde{Y}$ is the predicted outcome: which *risk category* has been assigned to the defendant?



Equal False Positive and False Negative Rates

$< \tilde{Y} \perp A \mid Y >$

The height proportion of the areas in color across same columns in each group should be the same

False

[Image from https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/]

The COMPAS Algorithm Revisited

- ***However …***

A few theoretical results about Fairness

Demographic Parity

$< \tilde{Y} \perp A >$

With a probabilistic predictor obtained from an historical dataset,
this can be attained only if $< Y \perp A >$, namely, if this is true in the records

But this is <u>not</u> a necessary condition for fairness (see Simpson's paradox)

Predictive Parity

$< Y \perp A \mid \tilde{Y} >$

Equal False Positive and False Negative Rates

$< \tilde{Y} \perp A \mid Y >$

These two conditions are mathematically *incompatible*, unless $< Y \perp A >$

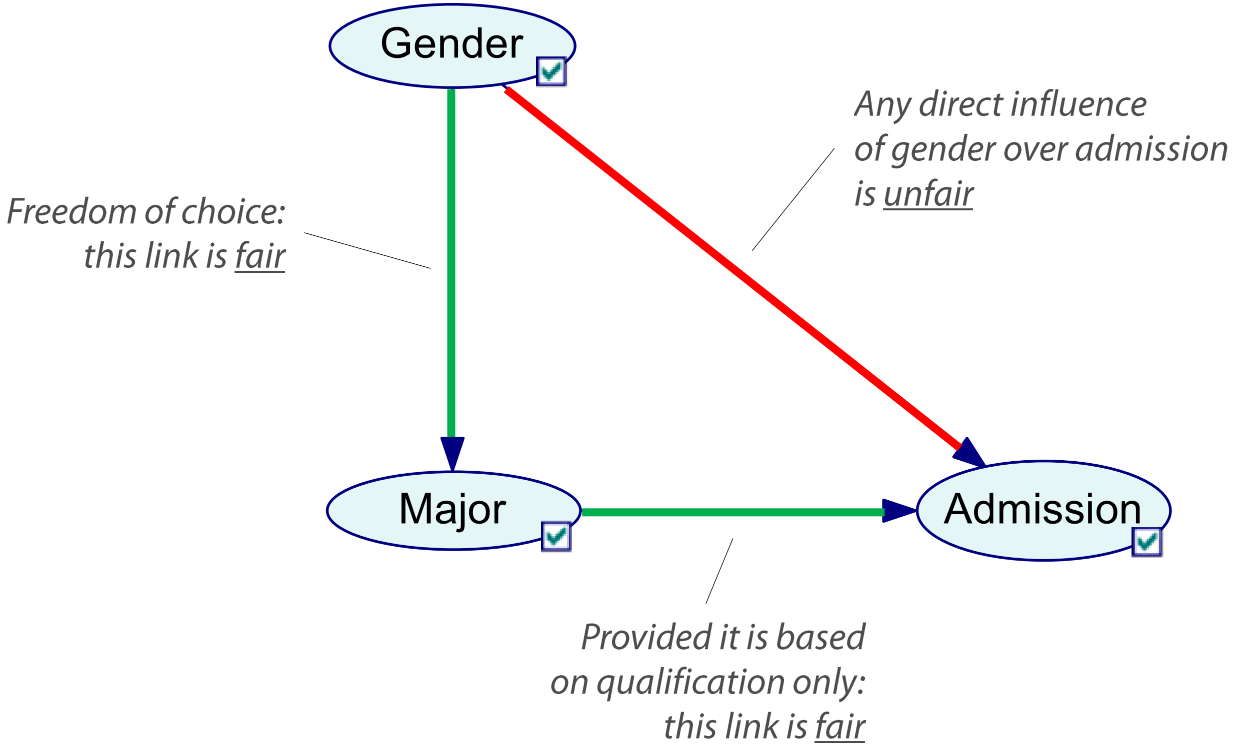
Once again, the latter is not a necessary condition <u>per se</u>

Causes and Effects, then Fairness

Path-Specific Effect

- ***Apropos Fairness***
  - Evaluating causes and effects, alone, may lead to paradoxical results
  - What 'stands in between' (*mediators* and *confounders*) needs to be considered as well

Causal Effects can be Path-Specific



Path-Specific Counterfactual Fairness

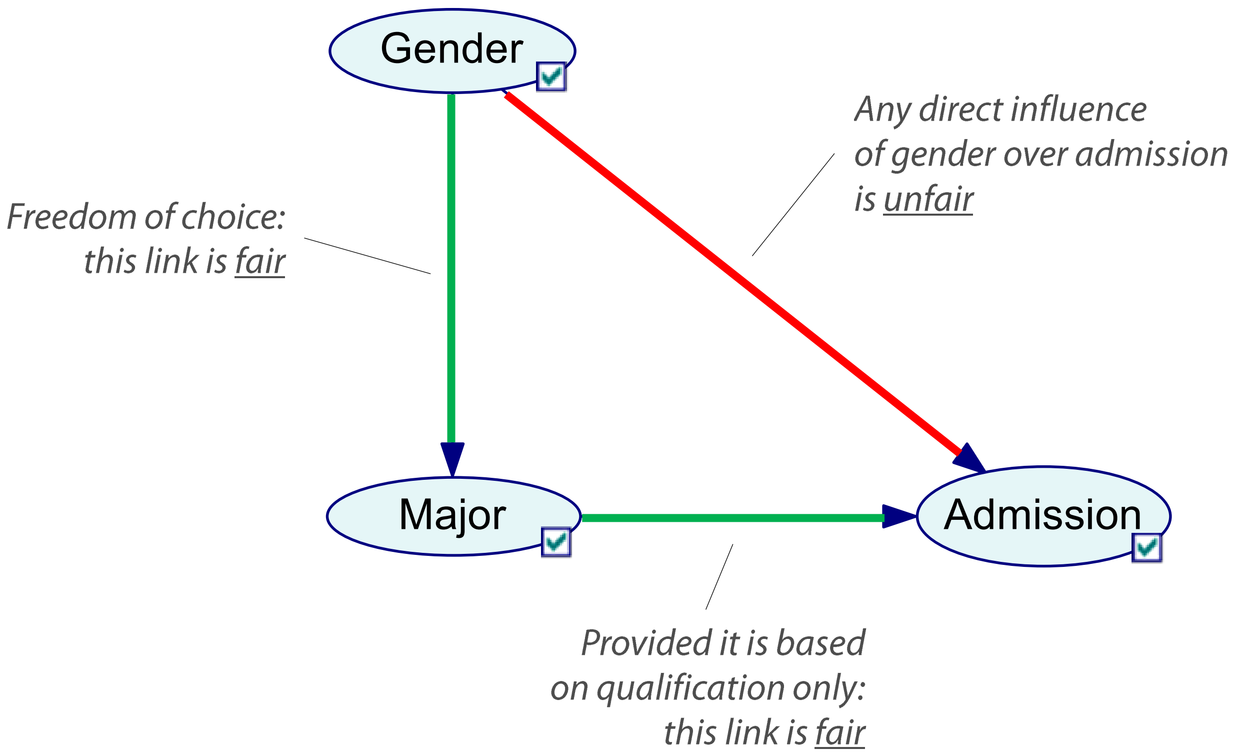
- ***Separating Effects: Counterfactuals***

  Ideally, we should estimate the causal effects along each path

  Or else....

Path-Specific Counterfactuals

What if a subject of one gender was of the other gender (*counterfactual*) along the unfair path?



Is this possible?
(see example in GeNIe)

Obtaining Fair Predictors

- ***Datasets are not necessarily fair***

  They contain historical data: no ‘a priori’ fairness assumptions

  We cannot change the past or just ignore sensitive attributes

  If applied blindly, *machine learning* will just reproduce the past

  Conundrum of AI predictors:

  - to make a predictor fair, we need to detach it from observations (i.e., the dataset)
  - then, what is the logical basis for prediction, at all?

Constrained optimization (for fairness)

  It is a sophisticated form of *machine learning* aiming to:

  - correct unfair information induced by the sensitive attribute
  - while retaining fair information

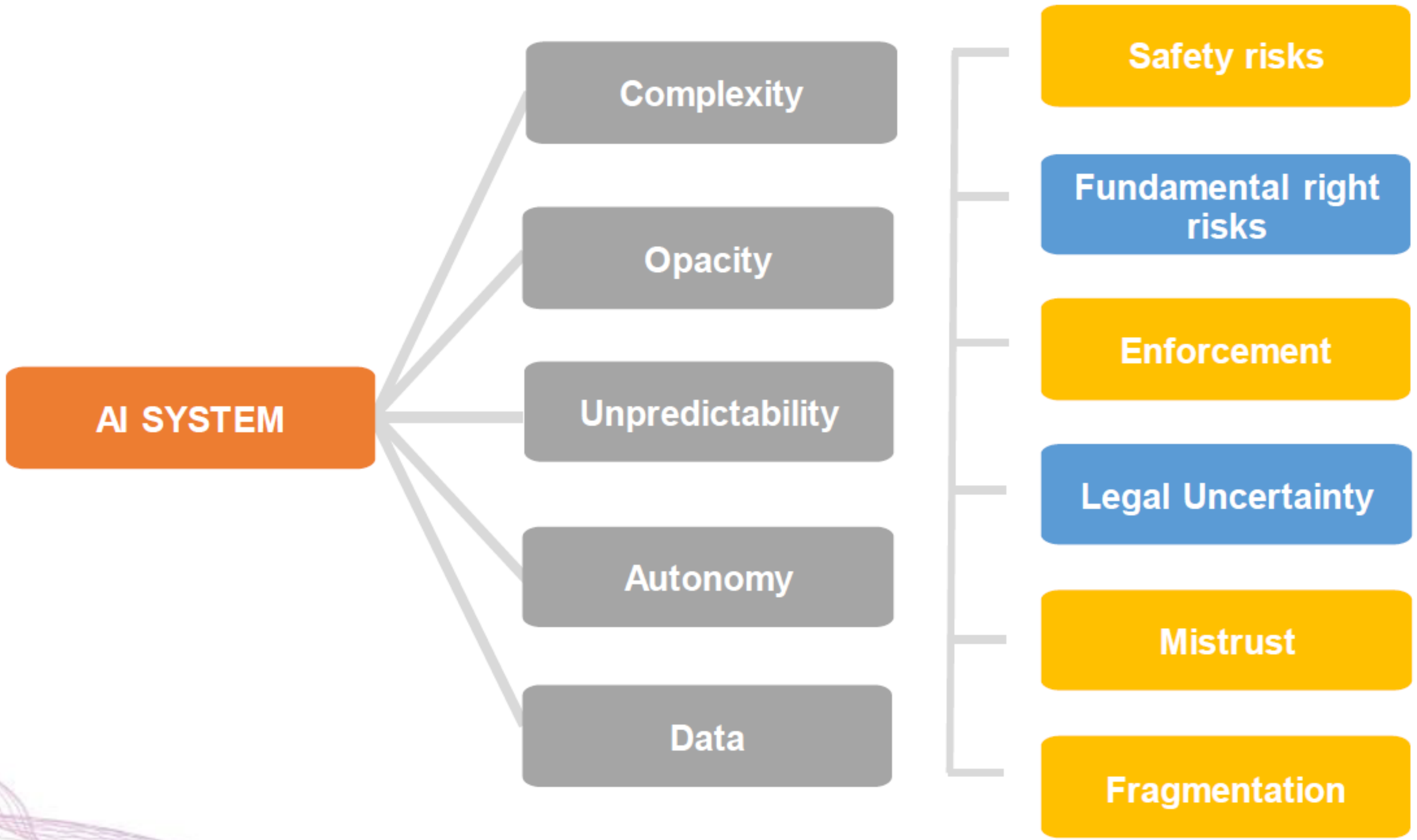
  It is a mathematically difficult technique, but feasible, with modern methods and machinery

  It is a very recent discipline (since around 2010)

An EU approach to the regulation of artificial intelligence

Artificial Intelligence Act [April 2021]

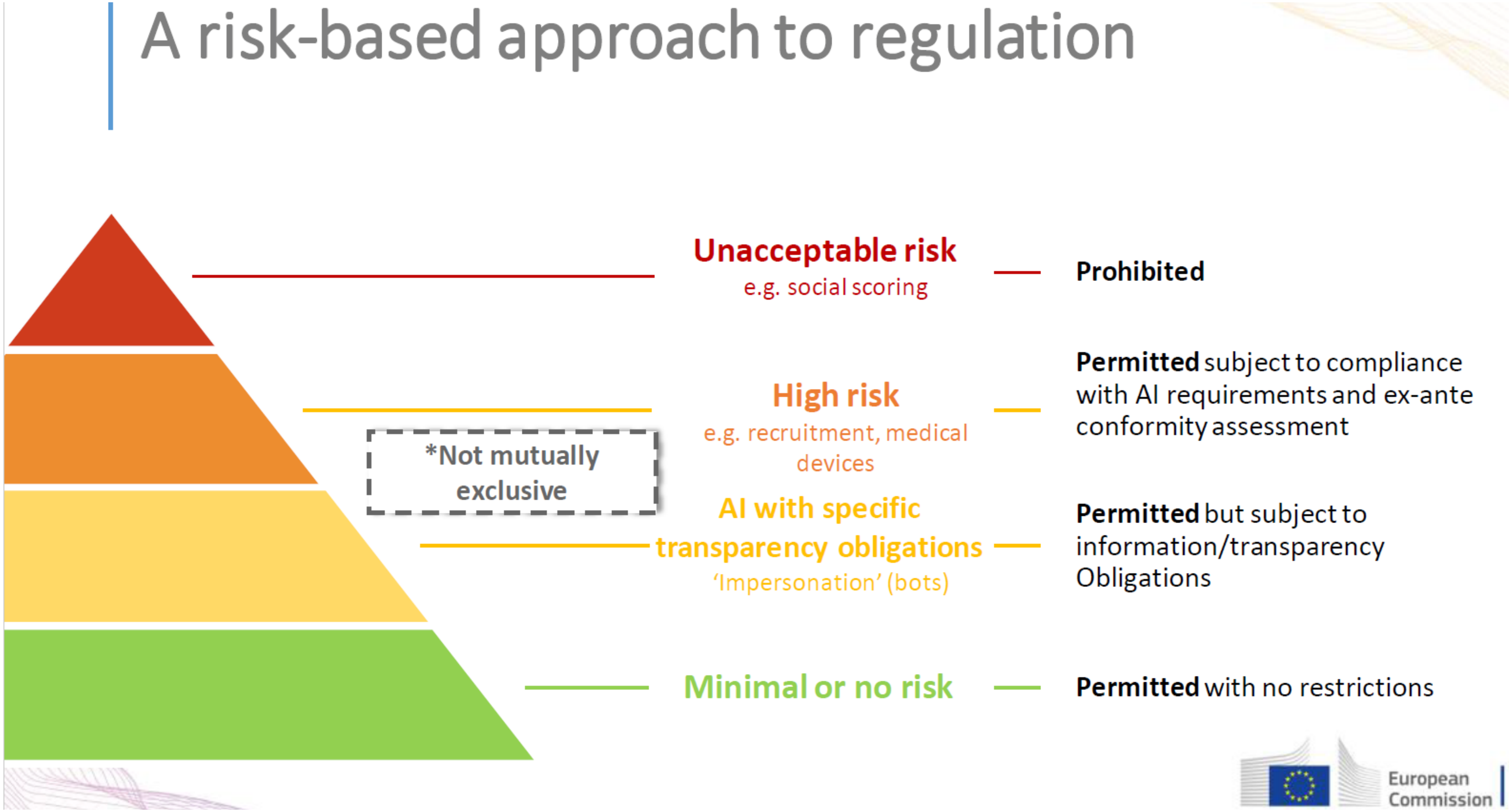
Why do we regulate AI use cases?



European Commission

[https://www.ceps.eu/wp-content/uploads/2021/04/AI-Presentation-CEPS-Webinar-L.-Sioli-23.4.21.pdf]

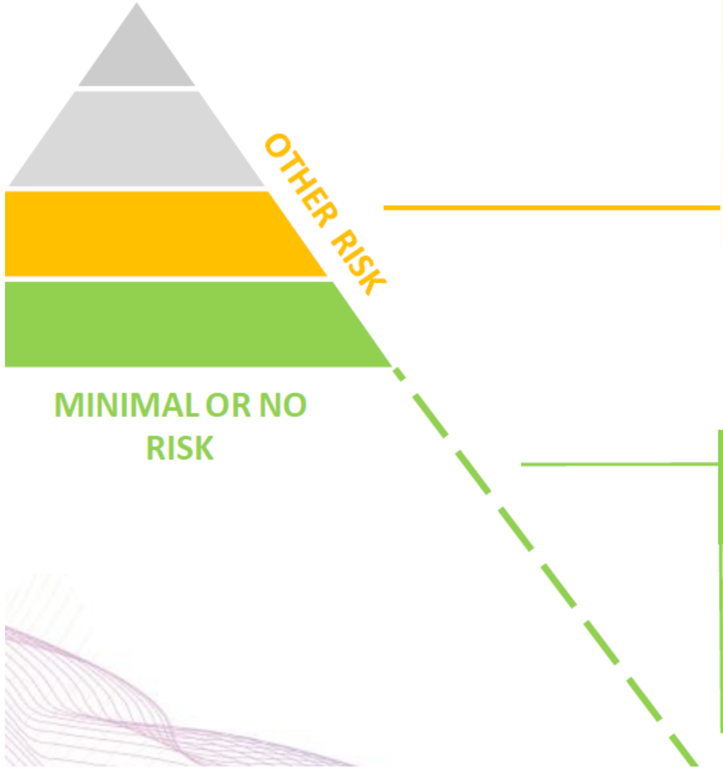
An EU approach to the regulation of AI



[https://www.ceps.eu/wp-content/uploads/2021/04/AI-Presentation-CEPS-Webinar-L.-Sioli-23.4.21.pdf]

An EU approach to the regulation of AI

Most AI systems will not be high-risk (Titles IV, IX)



New transparency obligations for certain AI systems (Art. 52)

- **Notify humans** that they are **interacting with an AI system** unless this is evident
- Notify humans that emotional recognition or biometric categorisation systems are applied to them
- Apply **label to deep fakes** (unless necessary for the exercise of a fundamental right or freedom or for reasons of public interests)

Possible voluntary codes of conduct for AI with specific transparency requirements (Art. 69)

- No mandatory obligations
- Commission and Board to encourage drawing up of codes of conduct intended to foster the **voluntary application of requirements to low-risk AI systems**

[https://www.ceps.eu/wp-content/uploads/2021/04/AI-Presentation-CEPS-Webinar-L.-Sioli-23.4.21.pdf]

High-risk Artificial Intelligence Systems (Title III, Annexes II and III)

Certain applications in the following fields:

1 SAFETY COMPONENTS OF REGULATED PRODUCTS

(e.g. medical devices, machinery) which are subject to third-party assessment under the relevant sectorial legislation

2 CERTAIN (STAND-ALONE) AI SYSTEMS IN THE FOLLOWING FIELDS

- ✓ Biometric identification and categorisation of natural persons
- ✓ Management and operation of critical infrastructure
- ✓ Education and vocational training
- ✓ Employment and workers management, access to self-employment
- ✓ Access to and enjoyment of essential private services and public services and benefits
- ✓ Law enforcement
- ✓ Migration, asylum and border control management
- ✓ Administration of justice and democratic processes

European Commission

[https://www.ceps.eu/wp-content/uploads/2021/04/AI-Presentation-CEPS-Webinar-L.-Sioli-23.4.21.pdf]

An EU approach to the regulation of AI

Requirements for high-risk AI (Title III, chapter 2)

Establish and implement **risk management** processes

&

In light of the **intended purpose** of the AI system

- Use high-quality **training, validation and testing data** (relevant, representative etc.)
- Establish **documentation** and design logging features (traceability & auditability)
- Ensure appropriate certain degree of **transparency** and provide users with **information** (on how to use the system)
- Ensure **human oversight** (measures built into the system and/or to be implemented by users)
- Ensure **robustness**, **accuracy** and **cybersecurity**

[https://www.ceps.eu/wp-content/uploads/2021/04/AI-Presentation-CEPS-Webinar-L.-Sioli-23.4.21.pdf]

An EU approach to the regulation of AI

Lifecycle of AI systems and relevant obligations

Design in line with requirements	Ensure AI systems **perform consistently for their intended purpose** and are **in compliance with the requirements** put forward in the Regulation
Conformity assessment	**Ex ante** conformity assessment
Post-market monitoring	Providers to **actively and systematically collect, document and analyse relevant data** on the reliability, performance and safety of AI systems throughout their lifetime, and to **evaluate continuous compliance of AI systems with the Regulation**
Incident report system	**Report serious incidents as well as malfunctioning leading to breaches to fundamental rights** (as a basis for investigations conducted by competent authorities).
New conformity assessment	**New conformity assessment** in case of **substantial modification** (modification to the intended purpose or change affecting compliance of the AI system with the Regulation) by providers or any third party, including when changes are **outside the "predefined range" indicated by the provider for continuously learning AI systems.**

[https://www.ceps.eu/wp-content/uploads/2021/04/AI-Presentation-CEPS-Webinar-L.-Sioli-23.4.21.pdf]

An EU approach to the regulation of AI

- *Where do we stand now*:

25 November 2022
The Council of the EU approved a compromise version of the proposed Artificial Intelligence Act

There are still disagreements in the definition of the AI systems.
The Council believes that the definition must not include certain types of existing software.
There are also difficulties in the definition of autonomy.

15 July 2022
Council of EU: Compromise text on the AI Act.

The Commission adopted the proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act, AIA) on 21 April 2021.

[Source https://www.artificial-intelligence-act.com/]

Thank you!