# Social and Ethical Implications of Artificial Intelligence

NELLO CRISTIANINI
PROFESSOR OF ARTIFICIAL INTELLIGENCE
UNIVERSITY OF BRISTOL

# The Idea

We have created **a version** of Artificial Intelligence, which is actually useful.

This was only made possible by a series of choices and shortcuts.

These shortcuts are also behind a series of problems.

Can we address these problems without losing the progress we made?

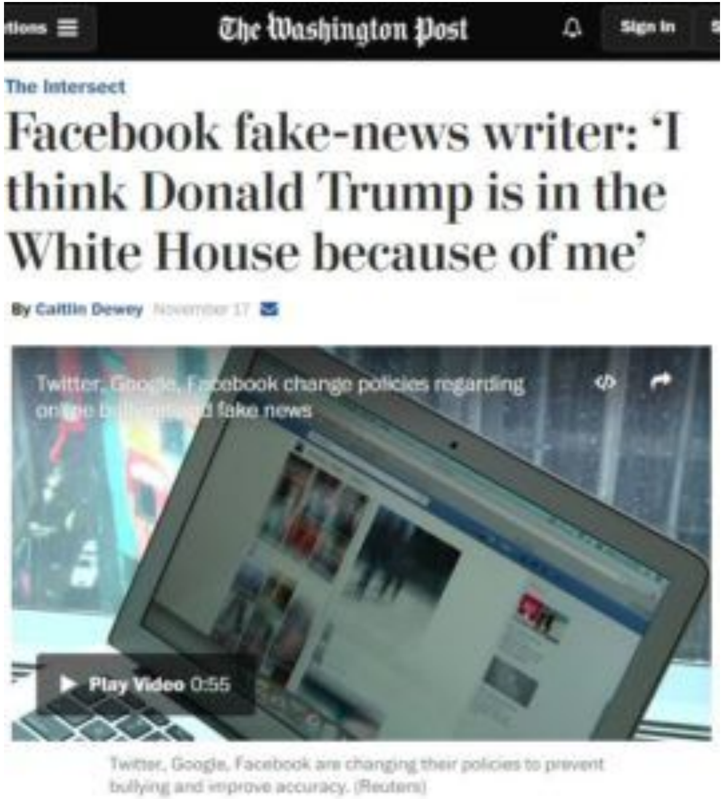
Either way, predictive accuracy is no longer sufficient to define the performance of AI agents. Transparency, Fairness, Privacy, User Autonomy, will all have to be baked in, somehow ...

# Steps

We have created intelligent agents with a series of <u>shortcuts</u> and cultural steps:

- Intelligence is about Behaviour (Turing, 1948)
- Prediction by statistics can replace modeling (eg: Halevy et al, 2009)
- Data from the wild, not handmade (eg: Halevy et al, 2009)
- Annotation from proxies, not direct (eg: **XXX** all early papers on implicit feedback)
- Social machines can be agents - humans can be participants (Cristianini et al, 2019)
- Agents can be intermediators, and media (Cristianini, 2018)
- ...
- Persuasion and Psychometrics (Burr et al, 2018 and 2019)

We deployed this version of AI at the centre of our infrastructure, it makes critical decisions, the consequences are becoming visible. It will not be simple to fix



The Washington Post

The Intersect

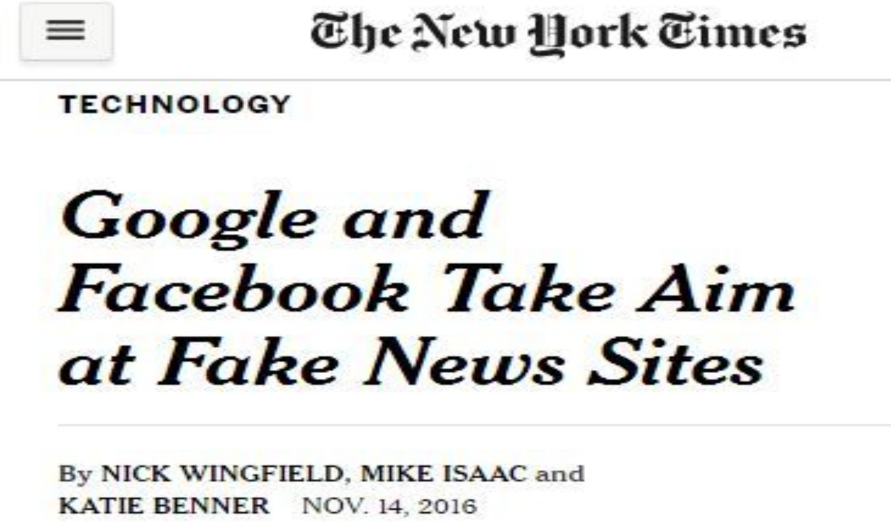
# Facebook fake-news writer: ‘I think Donald Trump is in the White House because of me’

By Caitlin Dewey November 17

Twitter, Google, Facebook change policies regarding online bullying and fake news

Twitter, Google, Facebook are changing their policies to prevent bullying and improve accuracy. (Reuters)

an OLD series of examples - but we do need to set up the context...



The New York Times

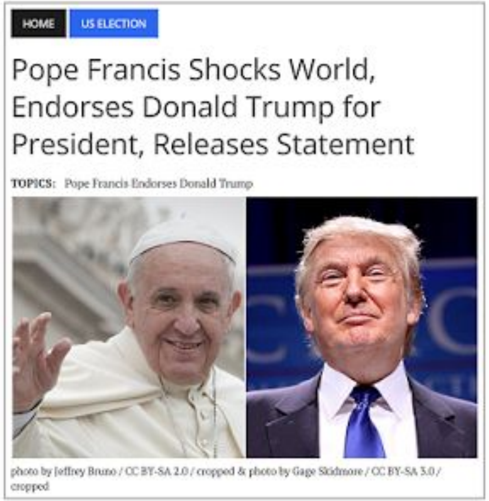
TECHNOLOGY

# *Google and Facebook Take Aim at Fake News Sites*

By NICK WINGFIELD, MIKE ISAAC and KATIE BENNER NOV. 14, 2016

# WTOE 5 NEWS

YOUR LOCAL NEWS NOW



HOME US ELECTION

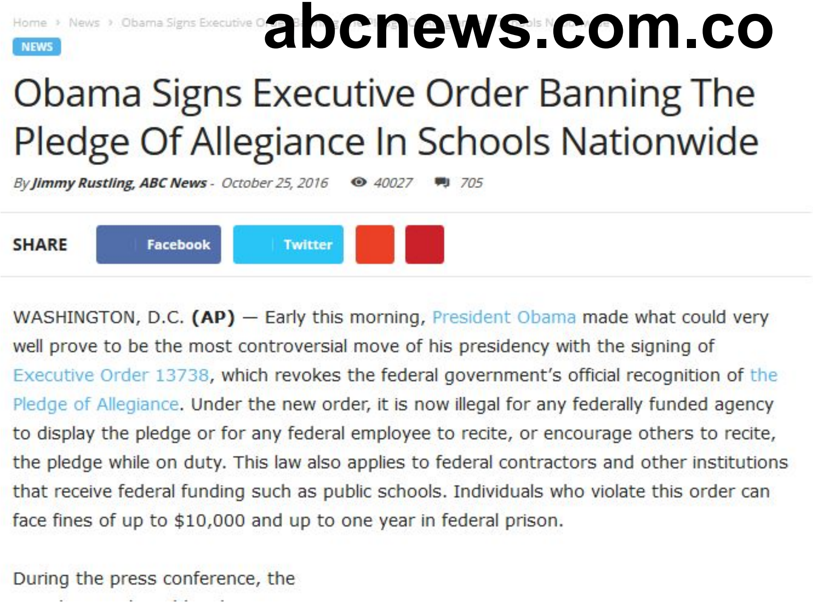
## Pope Francis Shocks World, Endorses Donald Trump for President, Releases Statement

**TOPICS:** Pope Francis Endorses Donald Trump

photo by Jeffrey Bruno / CC BY-SA 2.0 / cropped & photo by Gage Skidmore / CC BY-SA 3.0 / cropped

an OLD series of examples - but we do need to set up the context...

abcnews.com.co



Home › News › Obama Signs Executive O

NEWS

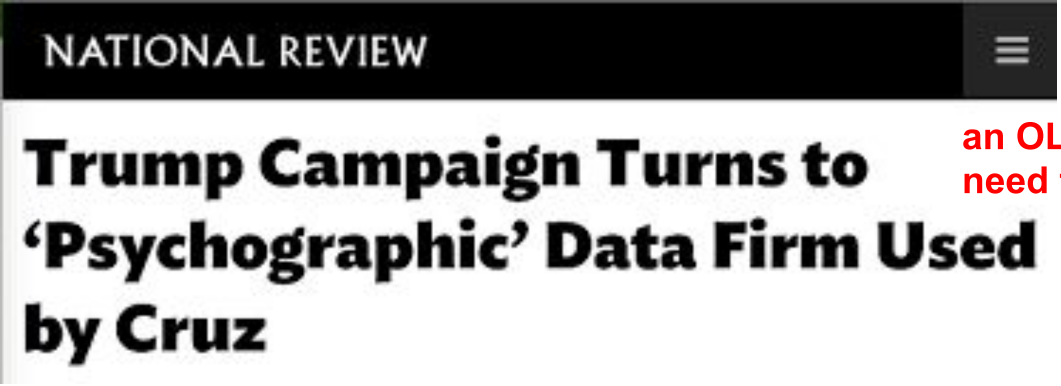
## Obama Signs Executive Order Banning The Pledge Of Allegiance In Schools Nationwide

By ***Jimmy Rustling, ABC News*** - *October 25, 2016* 👁 *40027* 💬 *705*

**SHARE** Facebook Twitter

WASHINGTON, D.C. **(AP)** — Early this morning, President Obama made what could very well prove to be the most controversial move of his presidency with the signing of Executive Order 13738, which revokes the federal government's official recognition of the Pledge of Allegiance. Under the new order, it is now illegal for any federally funded agency to display the pledge or for any federal employee to recite, or encourage others to recite, the pledge while on duty. This law also applies to federal contractors and other institutions that receive federal funding such as public schools. Individuals who violate this order can face fines of up to $10,000 and up to one year in federal prison.

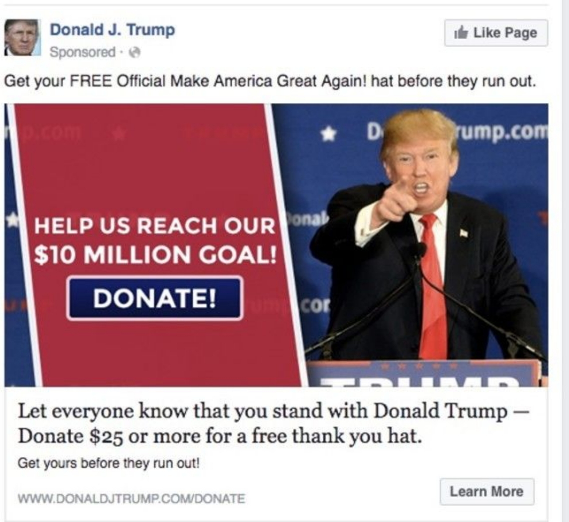
During the press conference, the



NATIONAL REVIEW

# Trump Campaign Turns to 'Psychographic' Data Firm Used by Cruz

an OLD series of examples - but we do need to set up the context...



**Donald J. Trump**
Sponsored ·

Like Page

Get your FREE Official Make America Great Again! hat before they run out.

HELP US REACH OUR $10 MILLION GOAL!

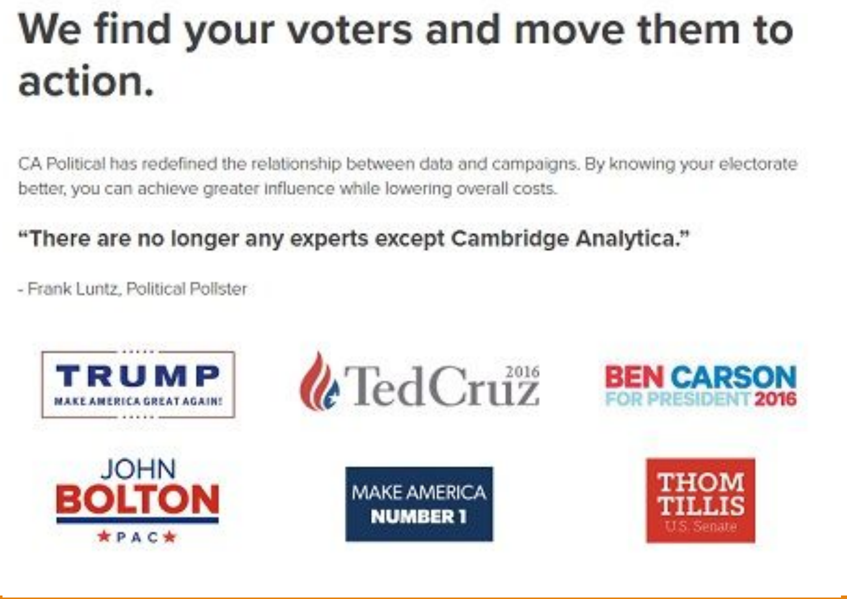
DONATE!

Let everyone know that you stand with Donald Trump — Donate $25 or more for a free thank you hat.

Get yours before they run out!

WWW.DONALDJTRUMP.COM/DONATE

Learn More



## We find your voters and move them to action.

CA Political has redefined the relationship between data and campaigns. By knowing your electorate better, you can achieve greater influence while lowering overall costs.

**"There are no longer any experts except Cambridge Analytica."**

- Frank Luntz, Political Pollster

TRUMP MAKE AMERICA GREAT AGAIN!

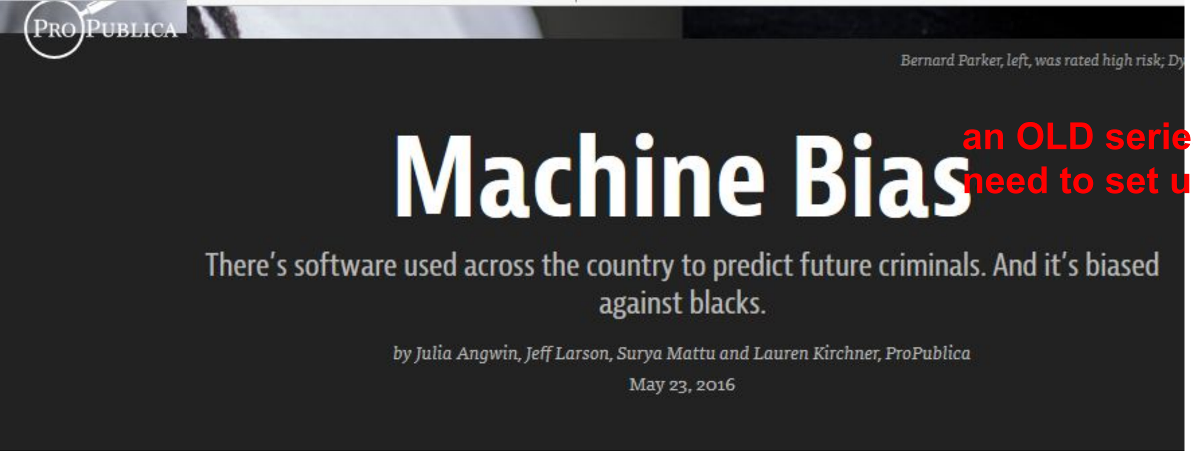
TedCruz 2016

BEN CARSON FOR PRESIDENT 2016

JOHN BOLTON ★PAC★

MAKE AMERICA NUMBER 1

THOM TILLIS U.S. Senate



ProPublica

Bernard Parker, left, was rated high risk; Dy

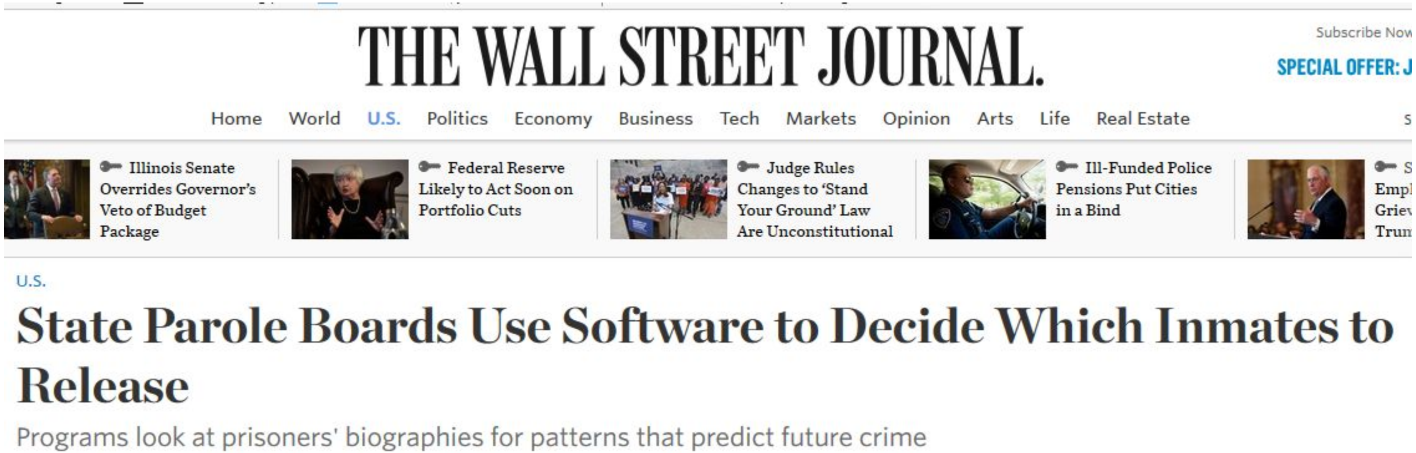
# Machine Bias

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

an OLD series of examples - but we do need to set up the context...



THE WALL STREET JOURNAL.

Subscribe Now

SPECIAL OFFER: J

Home World U.S. Politics Economy Business Tech Markets Opinion Arts Life Real Estate

Illinois Senate Overrides Governor's Veto of Budget Package

Federal Reserve Likely to Act Soon on Portfolio Cuts

Judge Rules Changes to 'Stand Your Ground' Law Are Unconstitutional

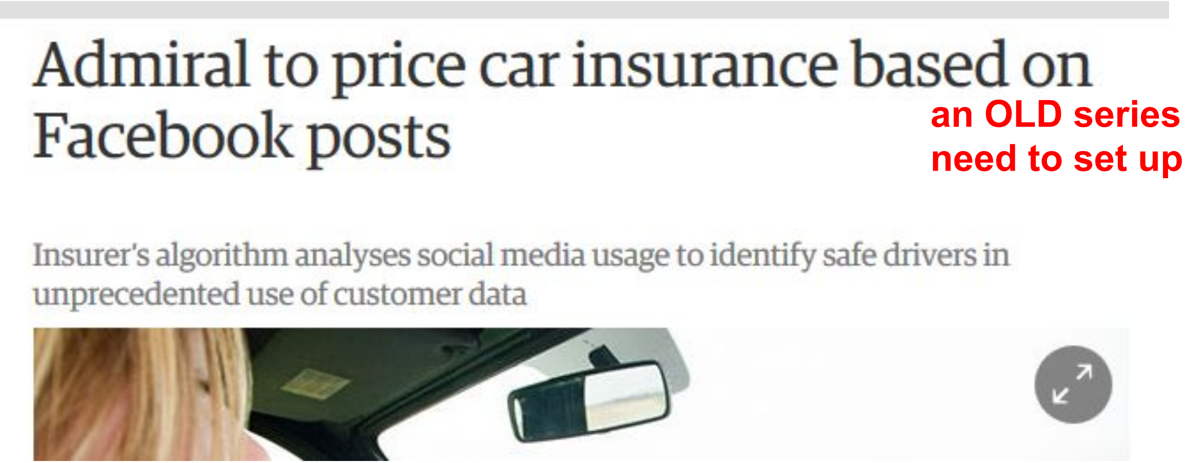
Ill-Funded Police Pensions Put Cities in a Bind

S Empl Griev Trun

U.S.

## State Parole Boards Use Software to Decide Which Inmates to Release

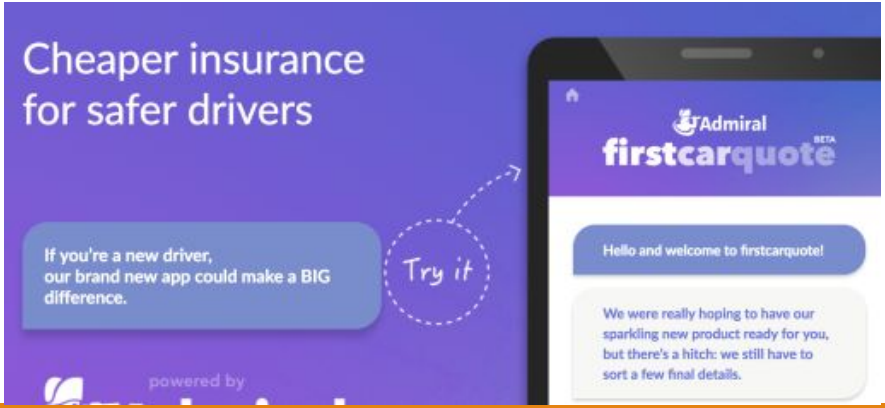
Programs look at prisoners' biographies for patterns that predict future crime



# Admiral to price car insurance based on Facebook posts

Insurer's algorithm analyses social media usage to identify safe drivers in unprecedented use of customer data

an OLD series of examples - but we do need to set up the context...



Cheaper insurance for safer drivers

If you're a new driver, our brand new app could make a BIG difference.

Hello and welcome to firstcarquote!

We were really hoping to have our sparkling new product ready for you, but there's a hitch: we still have to sort a few final details.

'I am the happy one' Mourinho's back – but has he really changed? By Daniel Taylor

Tuesday 11.06.13
Published in London
and Manchester
£1.40

theguardian
guardian.co.uk

# Europe demands answers from Obama over surveillance by US

an Travis
n Roberts Washington

k Obama was last night facing a
ing backlash in Europe against
eillance operations as leading
res demanded immediate
rom their American counter-
denounced the practice of
hering digital information on
s unacceptable, illegal and a
ion of basic rights.
kel indicated she would
on the revelations at a
ext week, while deputy
ission chief, Viviane

in American history. Hours after the video interview with Snowden appeared on the Guardian website, he checked out of the Mira Hotel in Hong Kong and moved to another, safer location in the region.

The Obama administration offered no indication of what it intended to do about Snowden, who was praised by privacy campaigners but condemned by some US politicians keen for him to be extradited and put on trial.

The White House did, however, say he had sparked an "appropriate debate" and hinted it might welcome revision of the Patriot Act, legislation introduced in 2001 which, it claims, gives legal authority

> “In my estimation, there has not been in American history a more important leak

an OLD series of examples - but we do need to set up the context...

Snooping scandal exclusive

UK used Google and Facebook to spy on us

theguardian

The whistleblower

I can't allow the US government to destroy privacy and basic liberties

theguardian

# What is happening?

Nothing new.

It is actually an old story…
… about new technology.

*It is about HOW we built this version of AI,*
*which assumptions and shortcuts we made.*
*It is a bit about the sociology of science too …*

# What is happening?

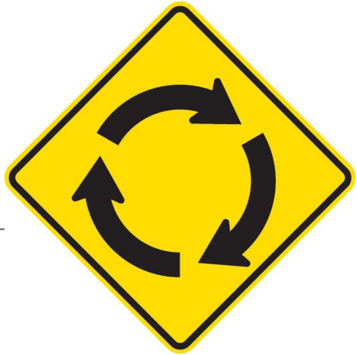
It is actually two stories...

Two technological breakthroughs have been converging fast for a while now…

A global Interconnected Data Infrastructure
Success in Artificial Intelligence (finally)

(trust me, this does matter)

# What is happening?



It is actually two stories...

Two technological breakthroughs have been converging fast for a while now…

A global Interconnected Data Infrastructure
Success in Artificial Intelligence (finally)

(trust me, this does matter)

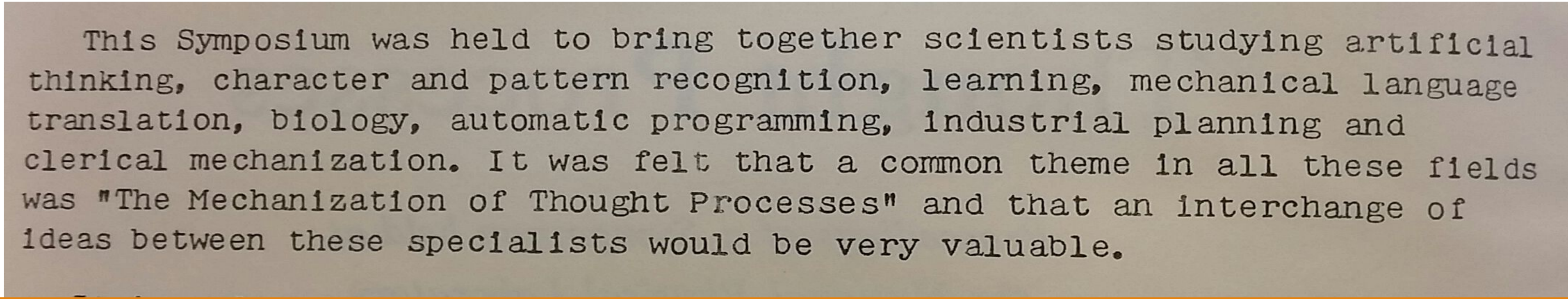
# Artificial Intelligence (before the 1990s)

Intelligent behaviour: expected to result from reasoning logically based on facts and rules that are known to the agent designer.

Examples: Chess games; Theorem provers.

# Artificial Intelligence (before the 1990s)

This approach **did not solve** any of the problems on the early lists of AI:
vision, translation, robot navigation, speech recognition,
question answering, robot navigation, etc



This Symposium was held to bring together scientists studying artificial thinking, character and pattern recognition, learning, mechanical language translation, biology, automatic programming, industrial planning and clerical mechanization. It was felt that a common theme in all these fields was "The Mechanization of Thought Processes" and that an interchange of ideas between these specialists would be very valuable.

November 1958

# Artificial Intelligence
# (the last 15 or 20 years)

The Unreasonable Effectiveness of Data

Alon Halevy, Peter Norvig, and Fernando Pereira, *Google*

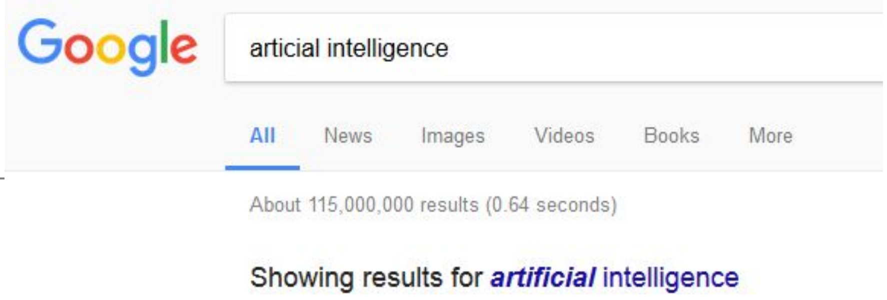
2009

1 - **Gather examples** about the behaviour to reproduce in the agent. (examples of translated text; transcribed speech; faces with names; correct spelling; discarded emails; … )

2 - Create **learning algorithms** that can learn from these examples, and reproduce the same behaviour.

**THE POINT:**

**this article captures a <u>mindset</u> that became prevalent in the early 2000s**



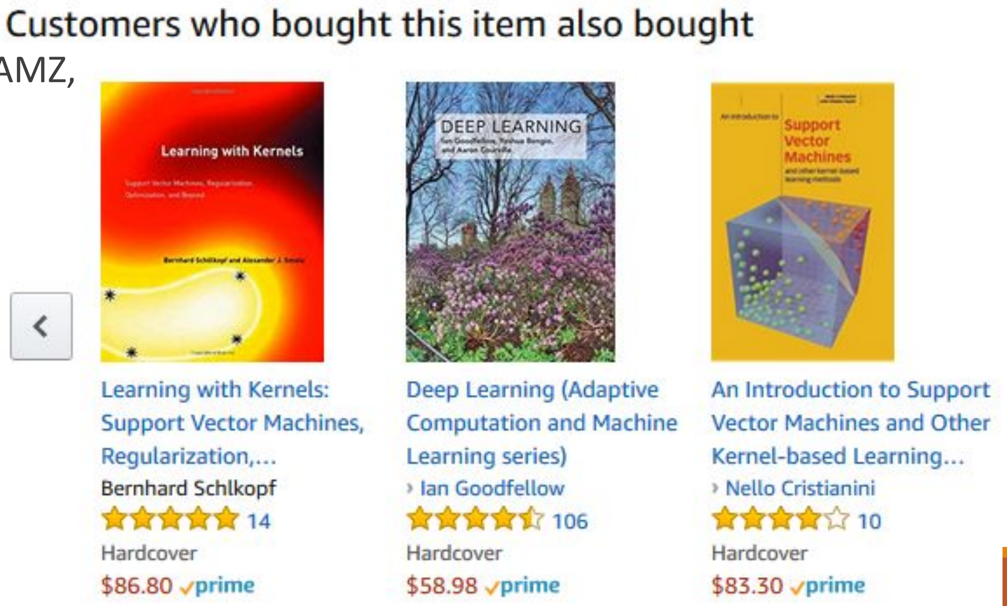
articial intelligence

About 115,000,000 results (0.64 seconds)

Showing results for *artificial* intelligence

# Artificial Intelligence

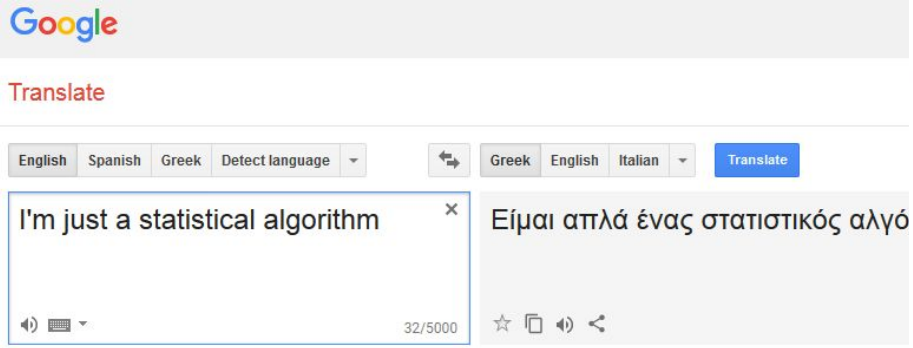
you can speak to Siri, translate with Google, recognise faces in FB, recommend books in AMZ, block spam in Hotmail, … …


Customers who bought this item also bought

Learning with Kernels: Support Vector Machines, Regularization,...
Bernhard Schlkopf
★★★★★ 14
Hardcover
$86.80 ✓prime

Deep Learning (Adaptive Computation and Machine Learning series)
› Ian Goodfellow
★★★★☆ 106
Hardcover
$58.98 ✓prime

An Introduction to Support Vector Machines and Other Kernel-based Learning...
› Nello Cristianini
★★★★☆ 10
Hardcover
$83.30 ✓prime



I'm just a statistical algorithm

32/5000

Είμαι απλά ένας στατιστικός αλγό

# ML+DATA=AI

THE ULTIMATE GO CHALLENGE
PAIR GO
26 MAY 2017
VS
Gu Li + AlphaGo
Lian Xiao + AlphaGo
RESULT W+Res

**Lady Lovelace was wrong**:
machines CAN do things
that their creators cannot understand
[my colleagues who worked on AlphaGo cannot defeat it]

Machine learning became the single recipe to solve AI problems.

“If an Intelligent Agent was a car, then the learning algorithm would be the engine, and the data would be the fuel… “

# The Age of Data

- The behaviour of the agent is driven by **correlations** discovered in the data by the learning algorithm - these are **not explanations** of the phenomenon - they just happen to work
- Discovery: many valuable behaviours can be emulated (and predicted) simply based on indirect correlations (across people, or different data sources)
- Unexpected data sources can be used to extract ‘signals’
- A popular book (*) summarised this as: 1) we do not need exactitude;  2) **correlation trumps causation**
- This created **a data rush** (“the new oil”)

(Cukier & Mayer-Schonberger)

**this also captures a mindset that became prevalent in the early 2000s — mindsets matter …**

# The power of Data

Data became
“the new oil” …

cisco

Data is the new oil.

We see in data the same transformative, wealth-creating power that 19th-century visionaries once sensed in the crude black ooze trapped underground.

The 2nd Annual European Data Economy Conference

BIG DATA

Towards a data-driven economy for Europe

25th March 2015 / Crowne Plaza Le Palace Hotel . Brussels

Hosted by The Software Alliance BSA

Organised by Forum europe

Partners DELL The power to do more Microsoft sas THE POWER TO KNOW

Delivering on the Digital Single Market

Building the European Data Economy

# The Shift

This shift is actually both part of the success, AND part of the new problems

(paradigm shift, sociology of science, but also the cause of today’s problems )

(see “ethical debt” article in publication)

# Getting the Data

How did AI agents manage to get access to the billions of examples that fuel every behaviour they need to produce?

How do they see enough discarded emails, shopping baskets, conversations, voices of different speakers, faces of different people, news reads, … ?

This is **another story**, that happened at the same time, and needs to be told…

**IT IS THE STORY OF HOW WE STARTED TO USE “DATA FROM THE WILD”**

# An Integrated Data Infrastructure ...

Before the AI revolution of the late 1990s, we used separate infrastructures to telephone, buy, bank, send letters, read news, …

But from the introduction of the WWW and its various layers, we gradually migrated most of that to the same new medium.

The same infrastructure and devices could now do payments, mail, shopping, news, entertainment, … (this was a massive migration)

How did that happen?

# On intermediaries, media, and middlemen ...

The pre-web world...

Who controlled access to the previous infrastructure?
How could a politician get elected, without the support of TV and newspapers?
How could a book be published, a song become famous, without the publishing industry?
How did you organise a trip without a travel agent?
How would you advertise your services?
Who set the prices and the conditions?

**The gatekeepers...**

# The big promise

An idea that made the new medium so revolutionary was the promise to **DIS-INTERMEDIATE** society, to directly connect politicians with voters, producers with customers, to bypass the traditional gatekeepers.

This would increase the power of the many, and decrease the power of the few.

Dis-intermediation as democratisation.

# The big promise

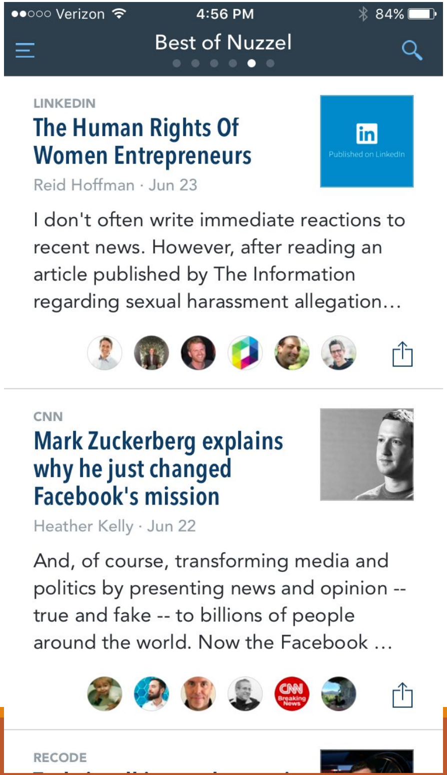
Autonomy for most people increased.

Favourite examples:

- videos of police brutality (would they have made it to TV ?)
- whistleblowing on corporations or politicians
- simpler ones: comparing prices, aggregating customer comments, making businesses compete

It worked. We adopted it .

# Selecting news, comparing prices, …



●●○○○ Verizon 4:56 PM 84%

Best of Nuzzel

LINKEDIN

**The Human Rights Of Women Entrepreneurs**

Reid Hoffman · Jun 23

I don't often write immediate reactions to recent news. However, after reading an article published by The Information regarding sexual harassment allegation…

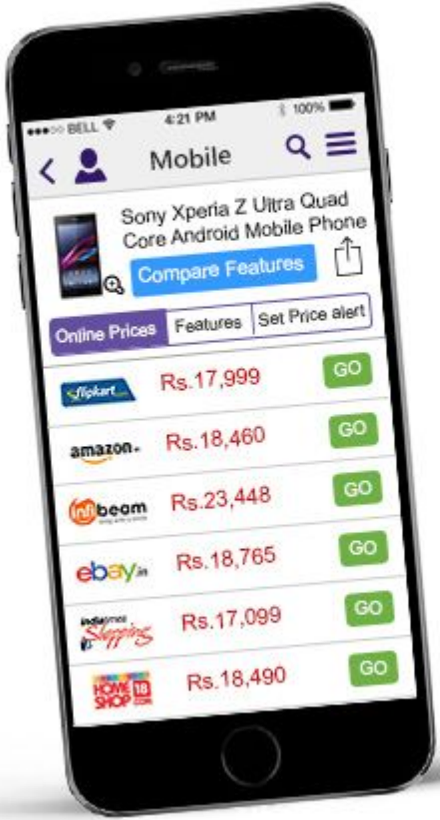
CNN

**Mark Zuckerberg explains why he just changed Facebook's mission**

Heather Kelly · Jun 22

And, of course, transforming media and politics by presenting news and opinion -- true and fake -- to billions of people around the world. Now the Facebook …

RECODE



●●●○○ BELL 4:21 PM 100%

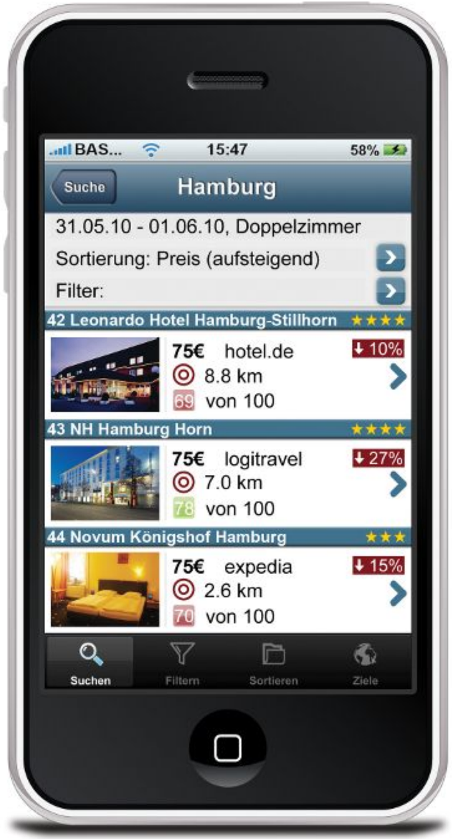
Mobile

Sony Xperia Z Ultra Quad Core Android Mobile Phone

Compare Features

Online Prices | Features | Set Price alert

| | | |
|---|---|---|
| flipkart | Rs.17,999 | GO |
| amazon. | Rs.18,460 | GO |
| infibeam | Rs.23,448 | GO |
| ebay.in | Rs.18,765 | GO |
| indiatimes Shopping | Rs.17,099 | GO |
| HOMESHOP18 | Rs.18,490 | GO |



BAS... 15:47 58%

Suche | Hamburg

31.05.10 - 01.06.10, Doppelzimmer

Sortierung: Preis (aufsteigend)

Filter:

42 Leonardo Hotel Hamburg-Stillhorn ★★★★
75€ hotel.de ↓10%
8.8 km
69 von 100

43 NH Hamburg Horn ★★★★
75€ logitravel ↓27%
7.0 km
78 von 100

44 Novum Königshof Hamburg ★★★
75€ expedia ↓15%
2.6 km
70 von 100

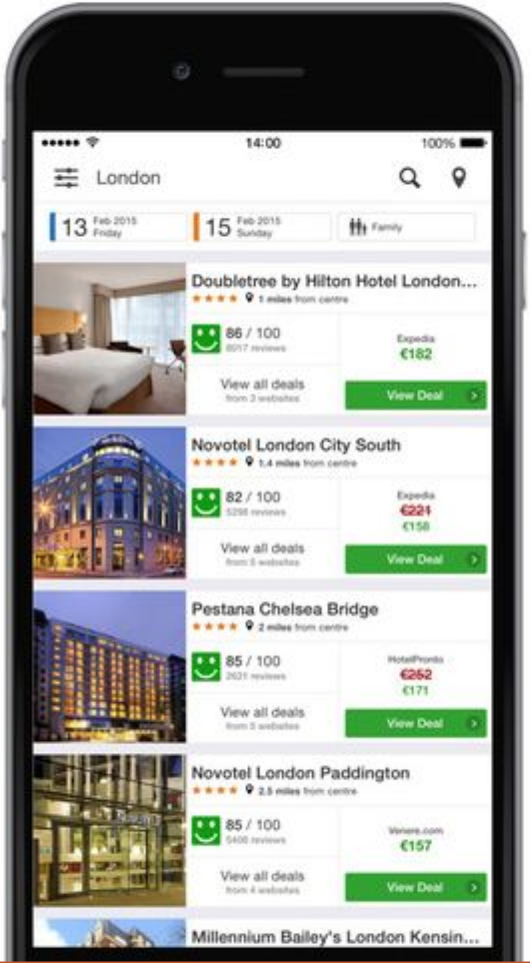
# … and collecting data.

At one end, delivering personalised intelligent services / suggestions / advice / connections / access (free to the customer, commission from the supplier).

At the other end (necessarily?) collecting data:

- who shopped for what, when and where?
- what did they actually buy?

This by-product of a sale may be as valuable as the commission.
WHAT IS THE ACTUAL BUSINESS MODEL ?



# Artificial Intelligence and Data Infrastructure - A Perfect Match...

As it grew the new infrastructure needed new algorithms to be usable, so that we can - find contents - screen money transactions - block spam - suggest new products ….

At the same time, as we used more of it, we generated vast masses of data for it to learn from: text, speech, purchases, actions, ratings, rankings, payments, etc …

Suddenly there was a new source of personal information about all of us, but there was also a new source of “fuel” for the machines...

The Unreasonable Effectiveness of Data

Alon Halevy, Peter Norvig, and Fernando Pereira, *Google*

# The “Unreasonable” paper spells out this “shortcut” ...

It identifies the causes for those successes in the availability of large amounts of data, already created for different purposes. “*In other words, a large training set of the input-output behaviour that we seek to automate is available to us in the wild. In contrast, traditional NLP problems such as (…) POS tagging (...) are not routine tasks so they have no large corpus available in the wild. Instead a corpus for these tasks requires skilled human annotation. Such annotation is not only slow and expensive to acquire, but also difficult for experts to agree on (...). The first lesson of web-scale learning is to use available data rather than hoping for annotated data which is not available. For example we find that useful semantic relationships can be learned from the statistics of web queries, or from the accumulated evidence of web-based text patterns and formatted tables, in both cases without needing any manually annotated data*”

# Implicit Annotation and Curation

Various methods were devised to ‘force’ the users to provide with the necessary annotation (eg supervision, labels, scores, ranking, etc)

This started early, for example: [Boyan et al, 1996]: “*we make a design decision not to require users to give explicit feedback on which hits were good and which were bad (… ) instead we simply record which hits people follow, (…) because the user gets to see a detailed abstract of each hit, we believe that the hits clicked by each user are highly likely to be relevant (… )*“.

# Implicit Feedback in retrieval AND RECOMMENDATION

The shift between retrieval and recommendation is very subtle, as they rely on the very same set of techniques. After being user in retrieval, *Implicit feedback* was also proposed as a way to improve recommender systems since 1998 [Oard and Kim, 1998], and clickthrough data were proposed since 2002 as a proxy for relevance in search engines [Joachims 2002]. From the late 1990s Amazon and others were making use of the feature "*people who bought this also bought …*" which also makes a clever use of implicit signals [Shafer et al, 1999].

# Data in the Wild - and AI

This data, so different from the artificial and synthetic testbeds
of past AI, came to be known as **DATA IN THE WILD**
- it already existed for its own reasons *in the wild*,
it would be treated *as a natural resource*.

Everyone was talking about data science, big data, and the new oil.
We received free services in return for data collection… why not?

- we replaced correlations for causations, statistics for models
- we replaced cheap gathered datasets for expensive crafted ones
- we replaced implicit feedback for expensive data curation
- ...

# Three Shortcuts (*)

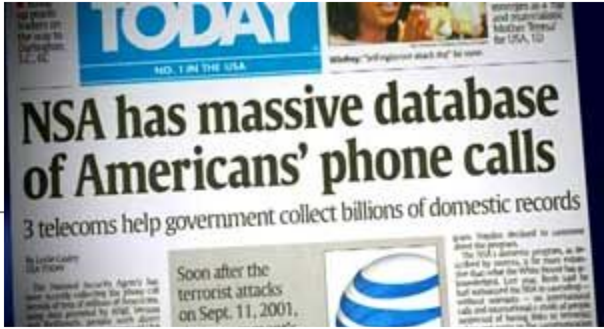
- From models to correlations … (non parametric models are not much better)
- From specifically generated data to data gathered from the wild
- From accurate annotation to proxies …

More shortcuts were taken as well, will discuss at the end

(*) forthcoming, (Cristianini 2019)

# Then one day we started seeing this...

And then the floodgates
opened and there was not going back ...

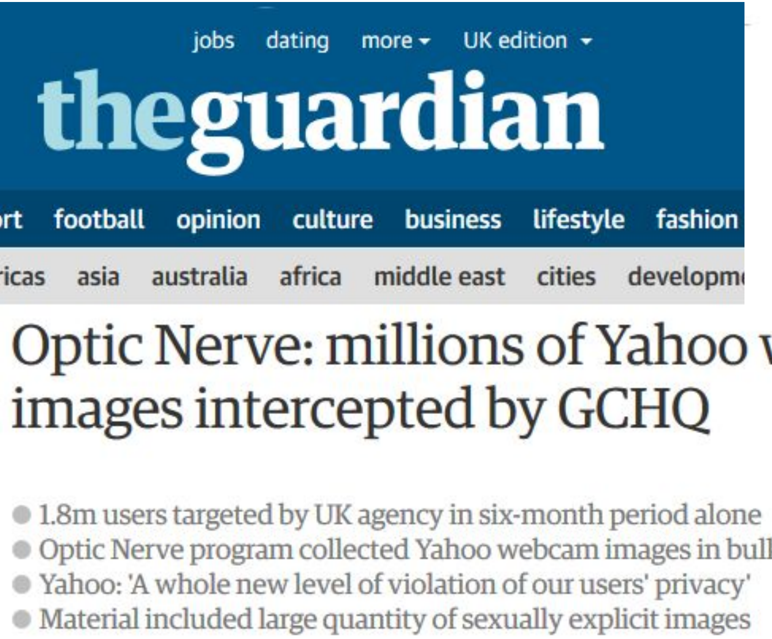

TODAY

NO. 1 IN THE USA

NSA has massive database of Americans' phone calls

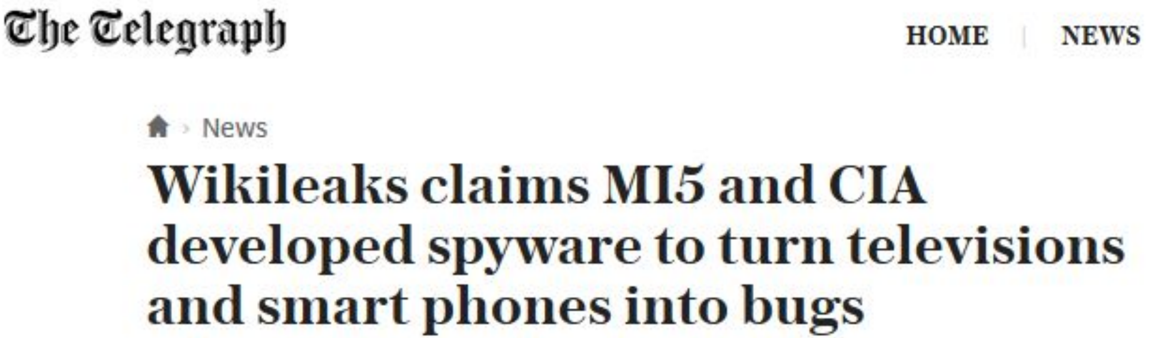
3 telecoms help government collect billions of domestic records

Soon after the terrorist attacks on Sept. 11, 2001,

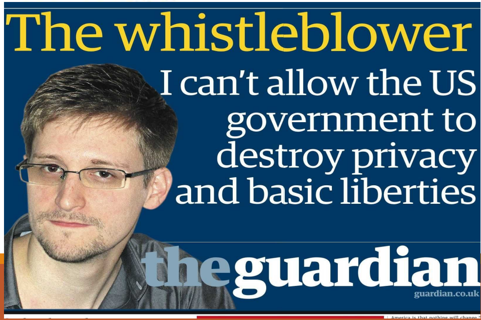


Optic Nerve: millions of Yahoo webcam images intercepted by GCHQ

- 1.8m users targeted by UK agency in six-month period alone
- Optic Nerve program collected Yahoo webcam images in bulk
- Yahoo: 'A whole new level of violation of our users' privacy'
- Material included large quantity of sexually explicit images

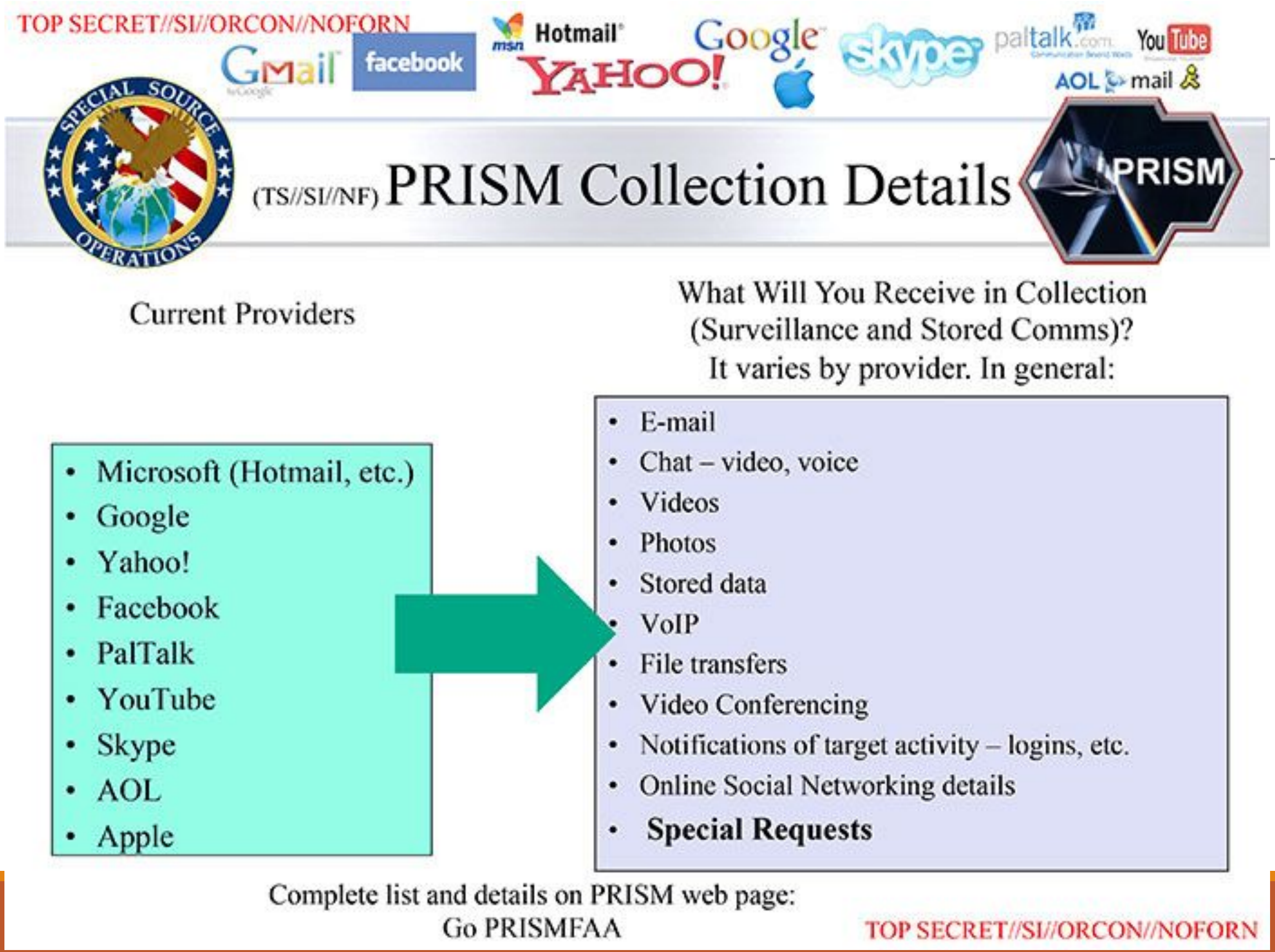


Wikileaks claims MI5 and CIA developed spyware to turn televisions and smart phones into bugs


The whistleblower

I can't allow the US government to destroy privacy and basic liberties

theguardian
guardian.co.uk



Gmail
facebook
Hotmail
Google
YAHOO!
skype
paltalk.com
YouTube
AOL mail
SPECIAL SOURCE OPERATIONS
PRISM

TOP SECRET//SI//ORCON//NOFORN

## (TS//SI//NF) PRISM Collection Details

Current Providers

- Microsoft (Hotmail, etc.)
- Google
- Yahoo!
- Facebook
- PalTalk
- YouTube
- Skype
- AOL
- Apple

What Will You Receive in Collection (Surveillance and Stored Comms)?
It varies by provider. In general:

- E-mail
- Chat – video, voice
- Videos
- Photos
- Stored data
- VoIP
- File transfers
- Video Conferencing
- Notifications of target activity – logins, etc.
- Online Social Networking details
- **Special Requests**

Complete list and details on PRISM web page:
Go PRISMFAA

TOP SECRET//SI//ORCON//NOFORN

**Many were surprised,**

**BUT…**

**...what else did we expect?**

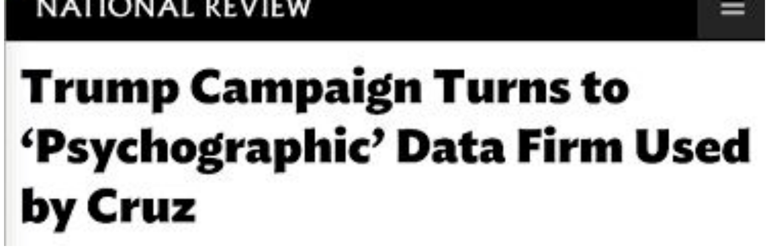
*https://www.theguardian.com/world/interactive/2013/nov/01/prism-slides-nsa-document*

# Cannot OPT OUT

We quickly realised that there was no opting out:
citizens can no longer function without being part of the system,
but the system can monitor their activities…

But mass surveillance and “passive” logging of online activities would soon be joined by new concerns…

# Targeting voters...



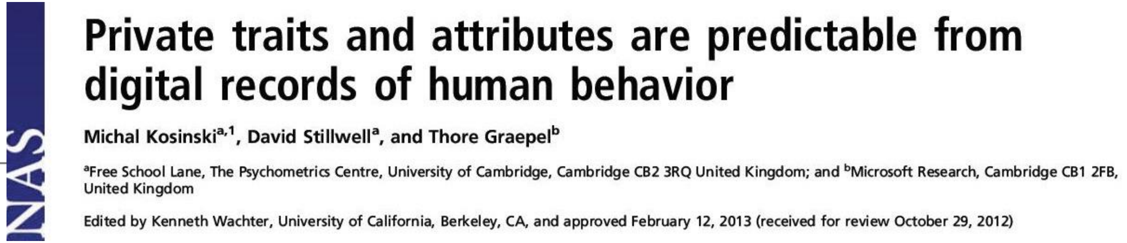
NATIONAL REVIEW

**Trump Campaign Turns to ‘Psychographic’ Data Firm Used by Cruz**

To Trump 2016 …
Cambridge Analytica (2016) added to the mix also 5 **personality traits** of voters, inferred from social media posts.
(ocean = openness, conscientiousness, extraversion, agreeableness, neuroticism)
(machine learning + personal data...)

[BUT: obama 2008 campaign not too different]

# What happened?



## 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

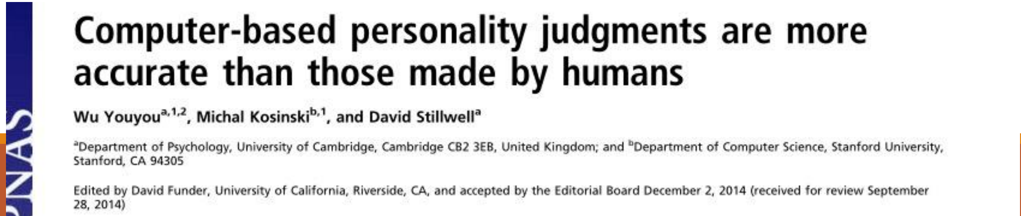
Edited by Kenneth Wachter, University of California, Berkeley, CA, and approved February 12, 2013 (received for review October 29, 2012)

Someone had created a dataset of hundreds of thousands of FB users, asking them to share their online content (posts, likes, friends, …) and to take a personality test, (and other psychological questionnaires) recording all the results. For many users it was a game.

Then they used machine learning to predict psychological features based on public FB activity.

Who would have thought? It worked.

(in the ‘not-exact’ + ‘not causal’ sense,
enough to place a bet)



## 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

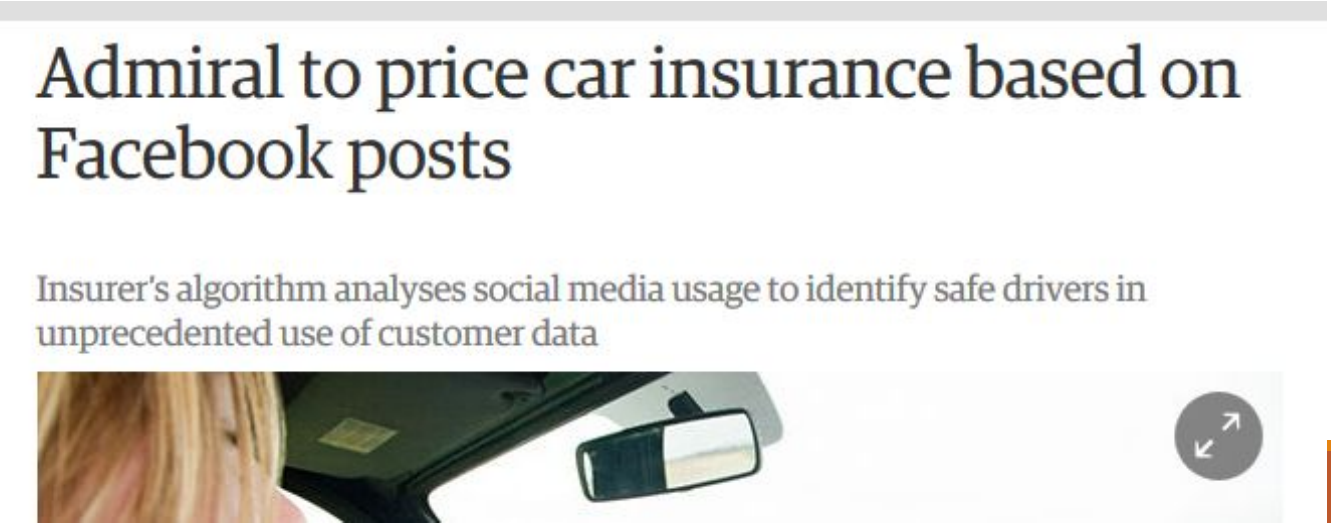
Edited by David Funder, University of California, Riverside, CA, and accepted by the Editorial Board December 2, 2014 (received for review September 28, 2014)

# Predicting Risk (or reducing opportunities?)

In 2016 Admiral Insurance proposed this … (and after media backlash withdrew it)

MACHINE LEARNING
+ PERSONAL DATA =

= psychometric
Information
⇒ assessing driving risk



Admiral to price car insurance based on Facebook posts

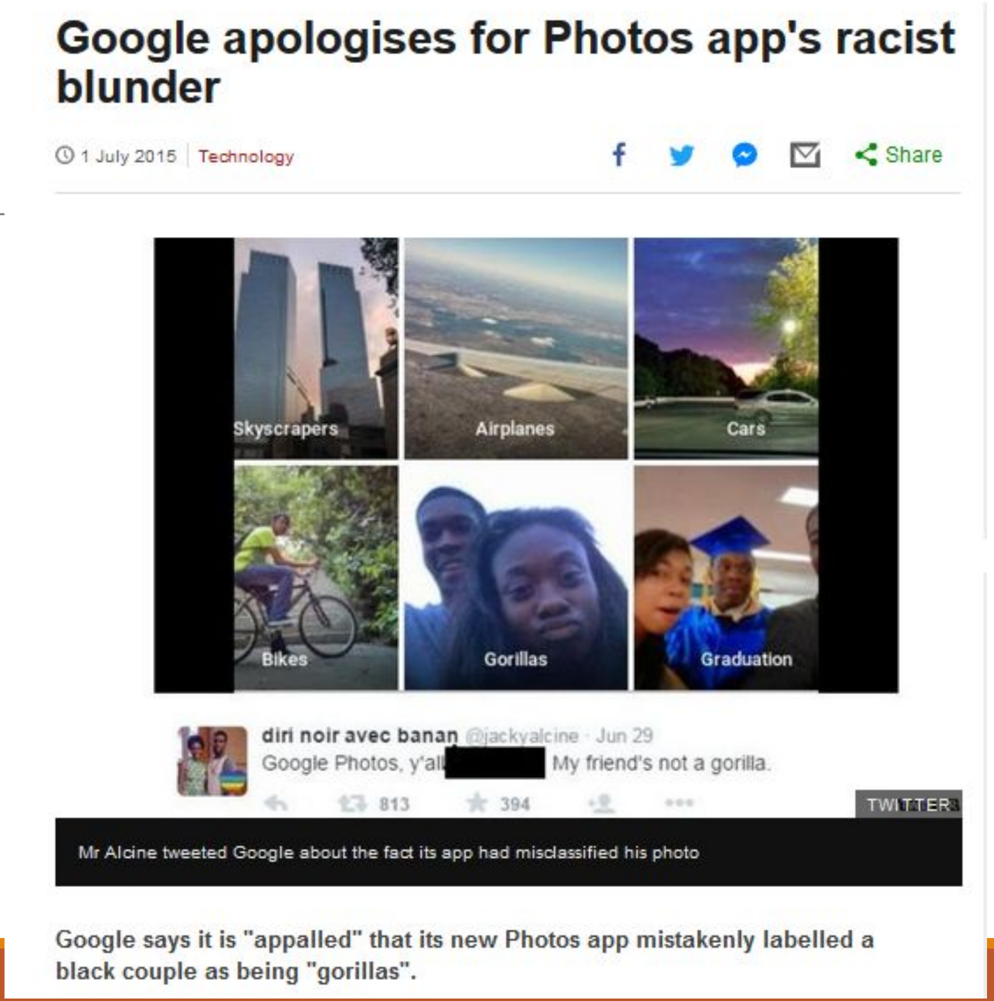
Insurer's algorithm analyses social media usage to identify safe drivers in unprecedented use of customer data

# Meanwhile in the News…

Nothing sinister or malicious…
… just another day in machine learning.

How were the training / testing set selected?
How was performance evaluated?
How was the tool eventually used?

Small statistical mismatches…
… nobody's fault.
Yet...



## Google apologises for Photos app's racist blunder

1 July 2015 | Technology

diri noir avec banan @jackyalcine · Jun 29
Google Photos, y'all [redacted] My friend's not a gorilla.

Mr Alcine tweeted Google about the fact its app had misclassified his photo

Google says it is "appalled" that its new Photos app mistakenly labelled a black couple as being "gorillas".

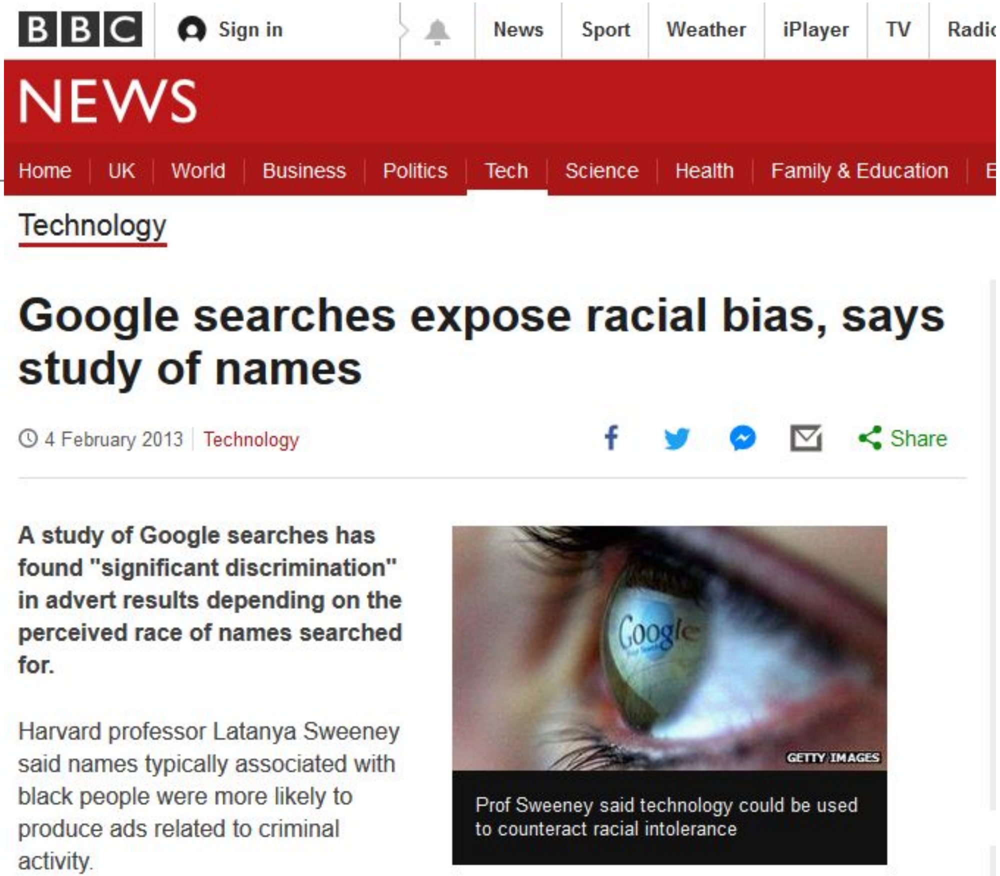
# Biased ads?

*“names linked with black people - were 25% more likely to have results that prompted the searcher to click on a link to search criminal record history.”*

SAME:...
How was training / testing done?

Data from the wild will reflect whatever biases already exist there… and AI will incorporate them into its agents.

Deep Learning algorithms cannot be inspected after training – we do not know what they have learnt.



## Google searches expose racial bias, says study of names

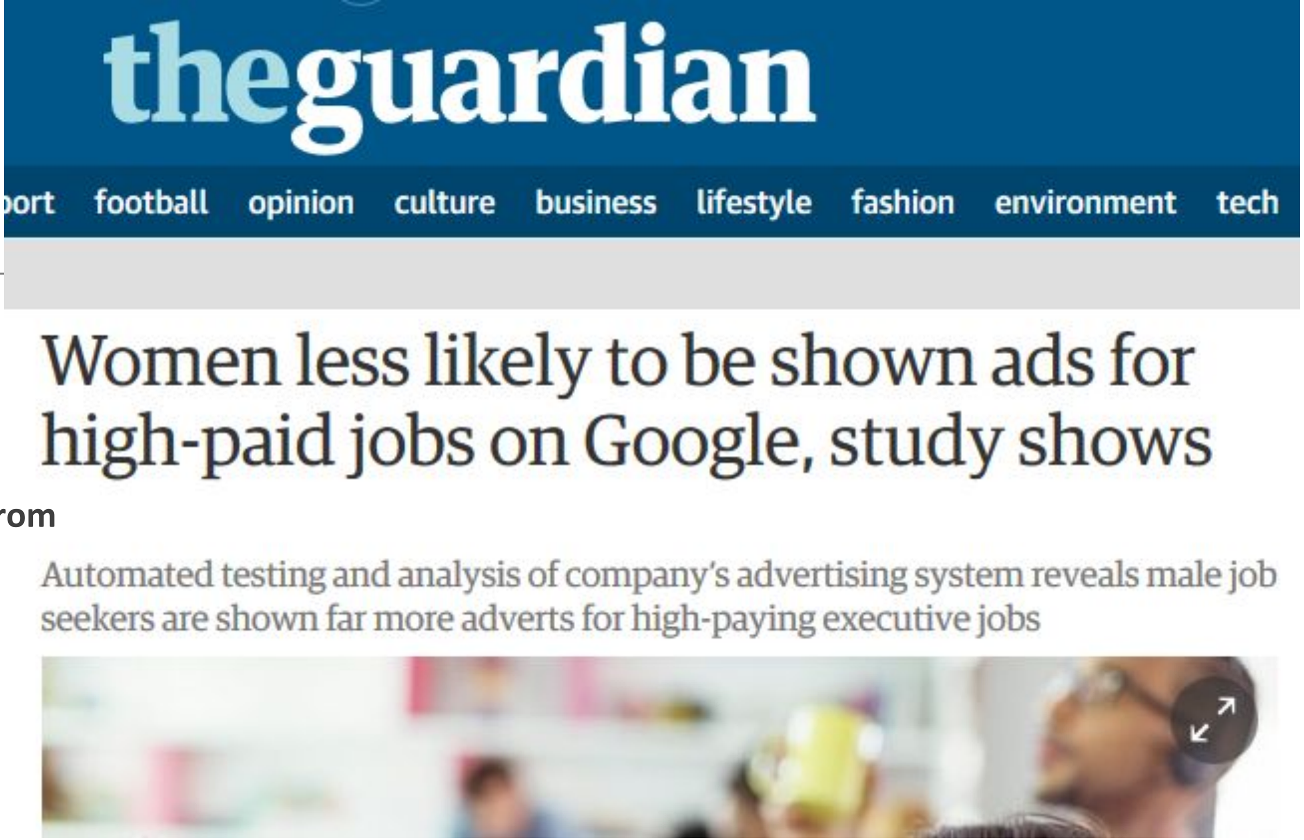
4 February 2013 | Technology

**A study of Google searches has found "significant discrimination" in advert results depending on the perceived race of names searched for.**

Harvard professor Latanya Sweeney said names typically associated with black people were more likely to produce ads related to criminal activity.

Prof Sweeney said technology could be used to counteract racial intolerance

# Biased ads? 2



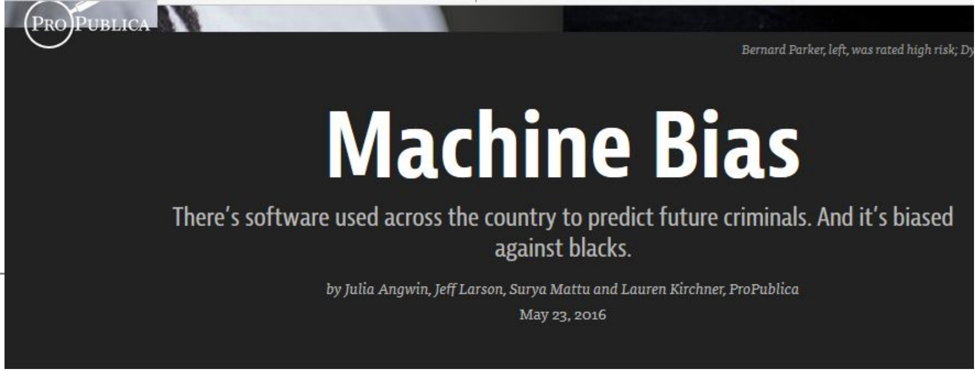
Women less likely to be shown ads for high-paid jobs on Google, study shows

Automated testing and analysis of company's advertising system reveals male job seekers are shown far more adverts for high-paying executive jobs

**Unexpected correlations**
**Are exactly what we want from ML algorithms…**
**… but here is one**
**we really did not expect !**

**Biases existing in the world end up in the training data… and finally in the AI…**
**(NOTE: My group has been analysing biases in media system and twitter for over ten years… )**

# Machine Justice, anyone?



Bernard Parker, left, was rated high risk; Dy

## 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

Various companies sell software to predict risk associated with releasing prisoners (under various conditions)...

E.g. : …

Features used include … **{ type of crime, age of first conviction, in some versions: address, education… }**

# Fake News

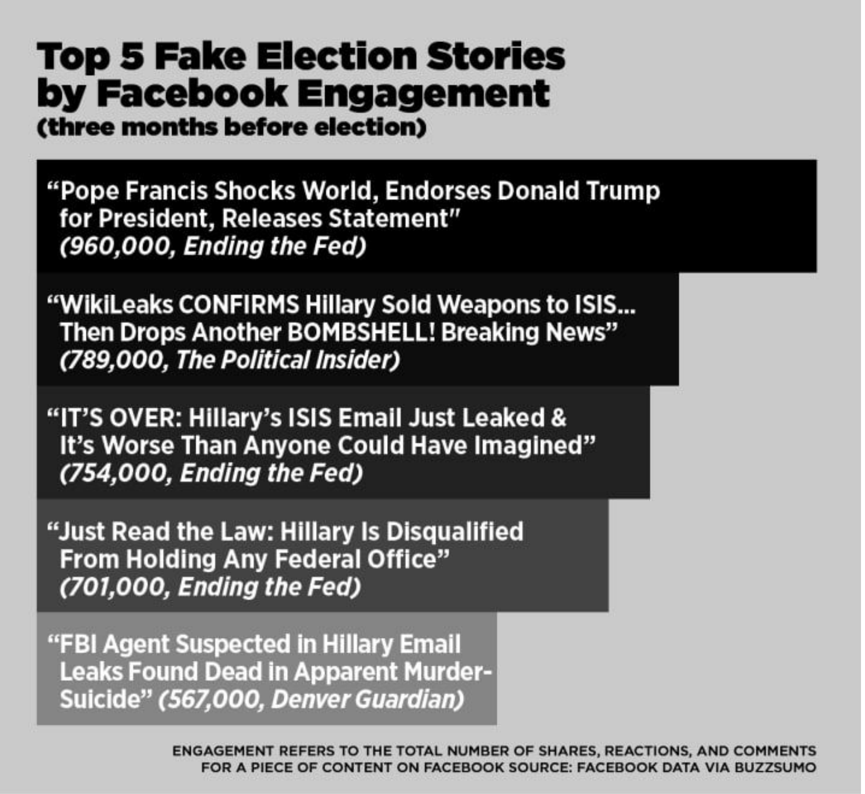
**Fake News <u>is an AI story</u>:**
Software recommends engaging stories on a personal basis

⇒ More of the same

⇒ The filter bubble / echo chamber

⇒ The **feedback loop**

⇒ Amplification / Polarisation



**Top 5 Fake Election Stories by Facebook Engagement**
**(three months before election)**

"Pope Francis Shocks World, Endorses Donald Trump for President, Releases Statement" *(960,000, Ending the Fed)*

"WikiLeaks CONFIRMS Hillary Sold Weapons to ISIS... Then Drops Another BOMBSHELL! Breaking News" *(789,000, The Political Insider)*

"IT'S OVER: Hillary's ISIS Email Just Leaked & It's Worse Than Anyone Could Have Imagined" *(754,000, Ending the Fed)*

"Just Read the Law: Hillary Is Disqualified From Holding Any Federal Office" *(701,000, Ending the Fed)*

"FBI Agent Suspected in Hillary Email Leaks Found Dead in Apparent Murder-Suicide" *(567,000, Denver Guardian)*

ENGAGEMENT REFERS TO THE TOTAL NUMBER OF SHARES, REACTIONS, AND COMMENTS FOR A PIECE OF CONTENT ON FACEBOOK SOURCE: FACEBOOK DATA VIA BUZZSUMO

Machine learning + personal data = personalised recommendation

FIGURE FROM: BUZZFEED.COM - https://www.buzzfeed.com/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook

# A New Medium

https://pixabay.com/en/monitor-eye-watch-spy-2276053/

We have **built a new mass medium**, that has no precedent in history, incorporating **AI**,

differently than telephone, telegraph, radio, TV,
this one is **aware of its contents (can stop, promote, translate, ...)**
**and looks back at us,** (and remembers)

we really need to use the theory of mass media...
(mcluhan, even popper, etc)

# AI, Generalisation and Personalisation

The data infrastructure can now recommend or block items just because they are sort-of similar to other items that people sort-of like us liked or disliked.

It can use this capability to generalise in order to *personalise* our relation with it

It is constantly refining its model of our individual preferences (such models exist multiple times within specific services, say online shops, search engines, or social networks, )

# So this is where we are...

-We have built a new medium, put at the centre of our lives
-It contains AI
-It looks at us, it learns from us
-It can learn the best of us, it can learn the worst of us
-We cannot opt out
-We have <u>not really removed intermediators, but replaced them </u>with intelligent algorithms ...

An interesting new problem for all of us...
⇒ **living with intelligent machines**.

# Living with Intelligent Machines

Earning Trust:

- **Fairness** (ensuring decisions are not biased, types of equality)
- **Transparency** (giving people control, explanations, black boxes, )
- **Accountability**
- **Privacy**

But also: besides regulating data collection, should we regulate the type of interventions that an AI machine can do (eg in the business of persuasion) ?

(*) Other issues exist, like employment, but today we focus on trust…

# GDPR and other opportunities

New EU Regulation, from May 2018, General Data Protection Regulation

GDPR can change a lot in the way we currently do AI (consent, transparency, right to erase, right to opt-out, ...)- (think about data brokers; cashless society; right to an explanation; ...)

That is the point of laws.

**The solution is not going to be only technical, but there is a place also for technical innovation.**

# Next challenges

**Algorithmic regulation** (Proposal to regulate society by means of algorithms to enforce laws…)

**Internet of Things**

**Cashless society**

- How would they work and what would be the benefits?
- More data collection
- Sounds like disintermediation all over again?
- Politicians: reasons for resisting pressures...

# A positive look - How do you earn trust?

How can we trust data-driven AI with the important parts of our lives (and culture, and public opinion, and the economy)

We need to understand the nature of the new problems: solutions will not come just from engineering or laws,... culture and politics need to evolve too.

- Explainable AI; Fairness by Design;
- GDPR + follow ups
- We might need new institutions (eg: like pharmaceutical industry, to license the use of algorithms?)
- Should monopolies be broken? (see portability in GDPR; competing for trust)

# Changes in the way we do machine learning ...

Success in AI and ML owes as much to social innovations among researchers as it does to technical innovations.

The specific variant of scientific method that we have evolved has been part of our success story.

Now it is time to take it to the next level …

# Cultural Steps - 1

The way we do AI today is due to an unexpected shift in the original path that was planned by the founders of the field: after decades of frustration, many of the key challenges in AI were conquered by adopting a pragmatic, data-driven approach.

This was made possible by Machine Learning and the availability of large amounts of data.

(see machine translation, computer vision, spam filtering, recommender systems...)

But: how did Machine Learning reach the point of being helpful in that pursuit?

This is because another cultural step that we should notice...

# Cultural Steps - 2

*The three shortcuts …*

- From models to correlations … (non parametric models are not much better)
- From specifically generated data to data gathered from the wild
- From accurate annotation to proxies …

*The use of social machines to generate intelligent agents … (humans as participants)*

*.*

# No free lunch?

- Machines can learn the meaning of words, and the behaviour of people, just by observing their everyday usage…
- Nobody can really check what they learn,
- As they absorb the meaning of words from subtle statistical relations in their usage … is it surprising that they also absorb our cultural biases?

# Next Cultural Steps...

***These shifts do matter - scientific progress is made of this sort of things***

***We probably need to move a little further now, predictive accuracy is no longer enough.***
***We are discovering that when we apply learning algorithms to people on this scale, performance is measured in a more complex way***

***Change is on its way…***

# The next dimensions of performance...

- **Fairness** - what does it actually mean? (agnostic embeddings and neural networks)
- **Readability / Explainability** (the right to an explanation in GDPR = meaning?)
- **Privacy**
- Etc

**Area in FAST expansion (21 definitions) - technical solutions are only PART of the solution - here are SOME examples from my group**

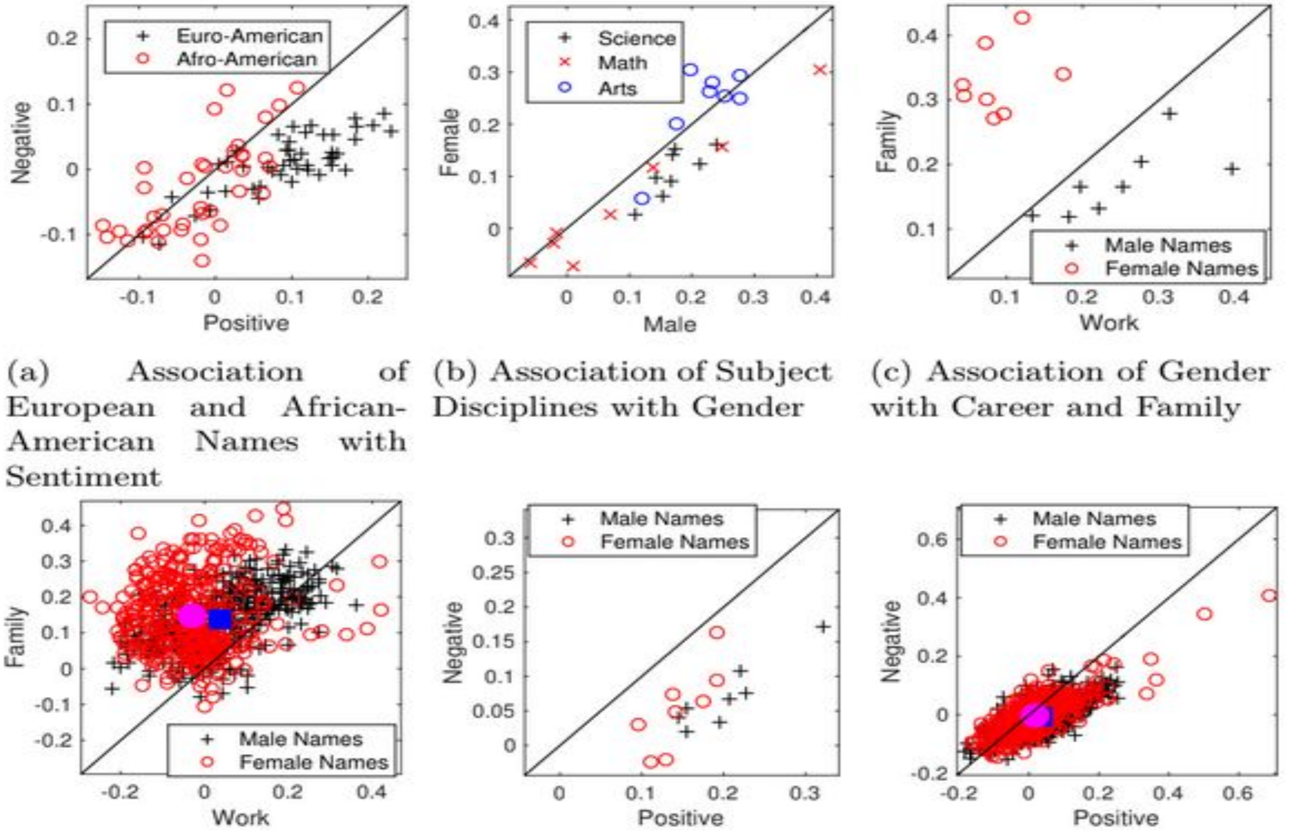
# Attempt 1 : WORD EMBEDDINGS

A very useful type of unsupervised learning …

It turns simple requirements / constraints into features … for words. We can finally capture semantic similarity between words, and use it in our inference …

It is part of the pipeline for most modern neural NLP systems

It is amazing - - - but has a problem. With potential discrimination ….



(a) Association of European and African-American Names with Sentiment

(b) Association of Subject Disciplines with Gender

(c) Association of Gender with Career and Family

(d) Extended Association of Gender with Career and Family

(e) Association of Gender with Sentiment

(f) Extended Association of Gender with Sentiment

Fig. 1: Association between different words and concepts in Experiment 1, resulting from the replication and extension of the Word Embedding Association Tests.

# Can this be fixed?

… our algorithm …

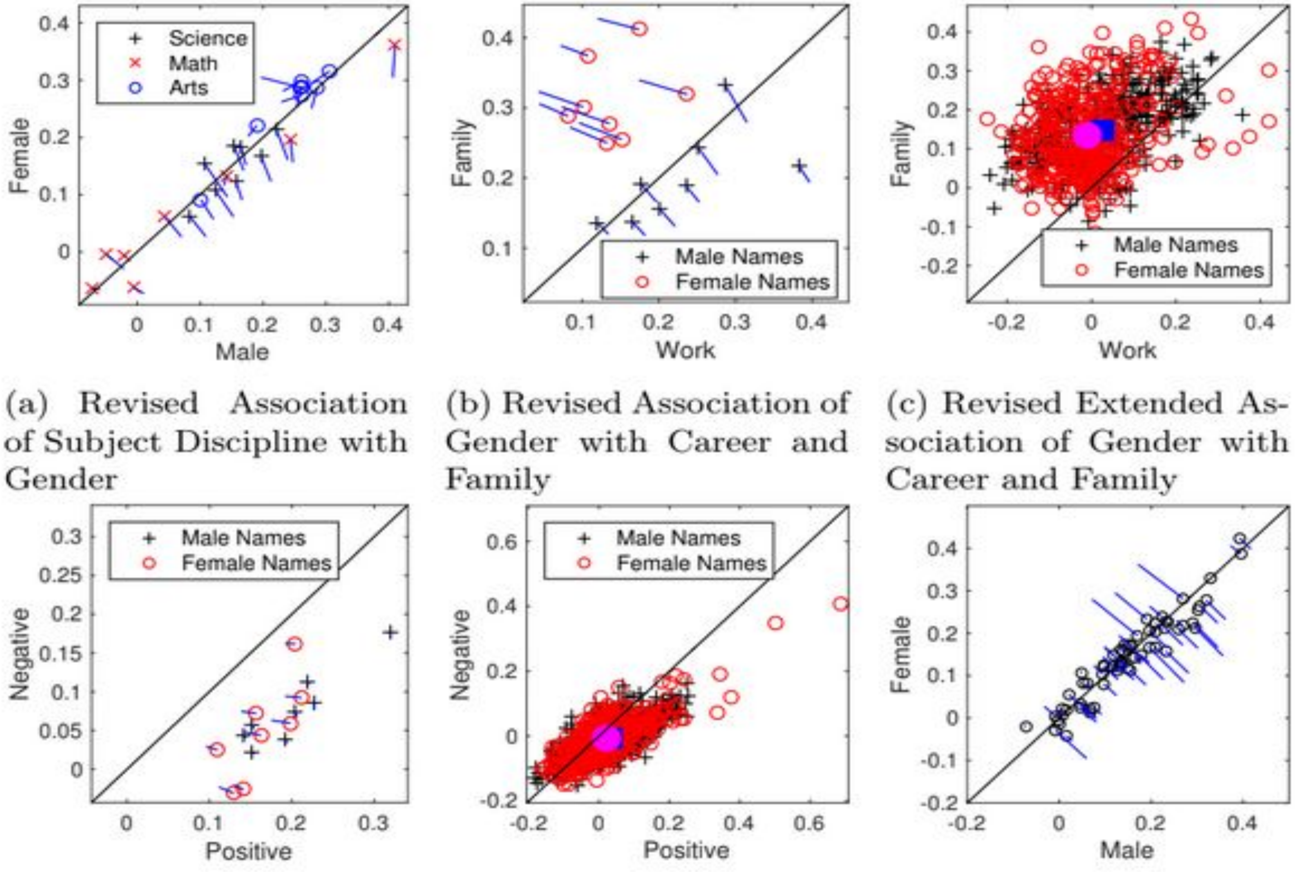
# Fairness

Equality of opportunity

Equality of outcome

How do we make sure that all candidates have the same opportunities, and are not judged based on forbidden dimensions?

(sometimes these are implicitly present in other data, like post codes, etc)



(a) Revised Association of Subject Discipline with Gender

(b) Revised Association of Gender with Career and Family

(c) Revised Extended Association of Gender with Career and Family

(d) Revised Association of Gender with Sentiment

(e) Revised Extended Association of Gender with Sentiment

(f) Revised Association of Occupation with Gender

# Attempt 2 - Right for the Right Reason...

… how do we know that a decision about a person is not made based on illegal bias?

(eg gender or race in job candidate selection, etc) … ?

A neural network could find a way to extract forbidden information based on legal features, and maybe use it - without the programmer knowing it

- Information leaks exist -

How do we rule out that forbidden information is used by a NN?

# Learning a target WITHOUT using a correlated background

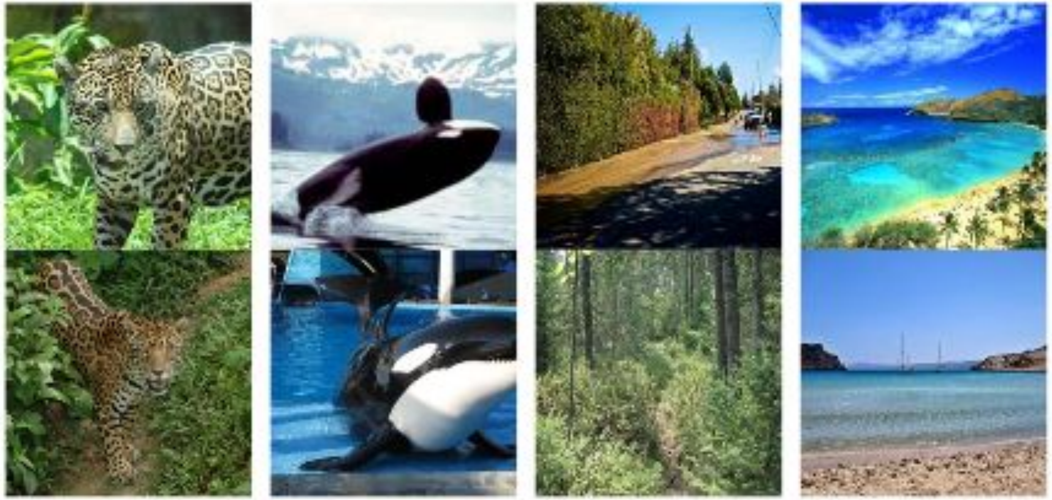
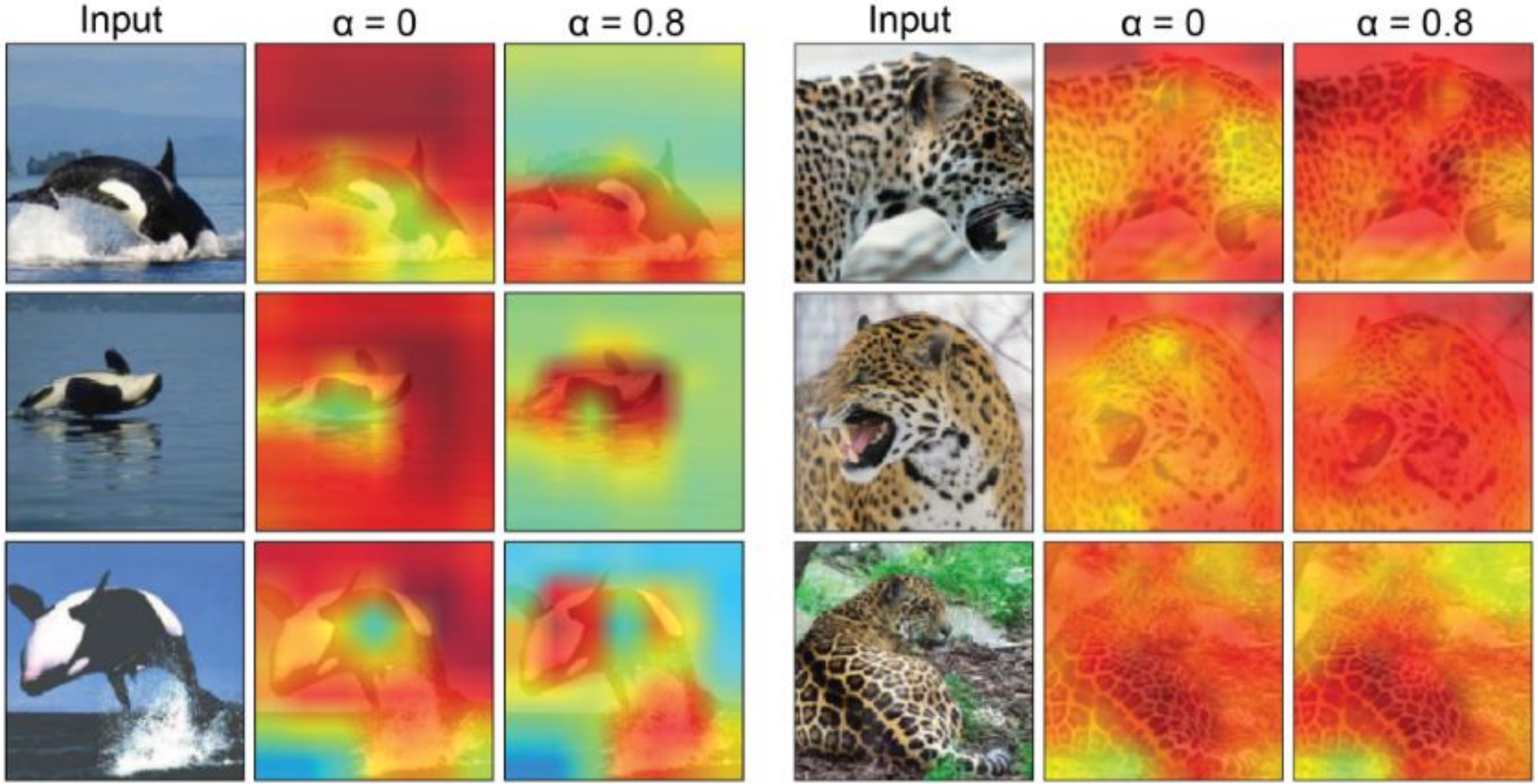


Fig. 1. Example images taken from the 'Jaguar', 'Killer whale', 'Forest path' and 'Coast' categories of the ImageNet and Places datasets respectively (left-right).

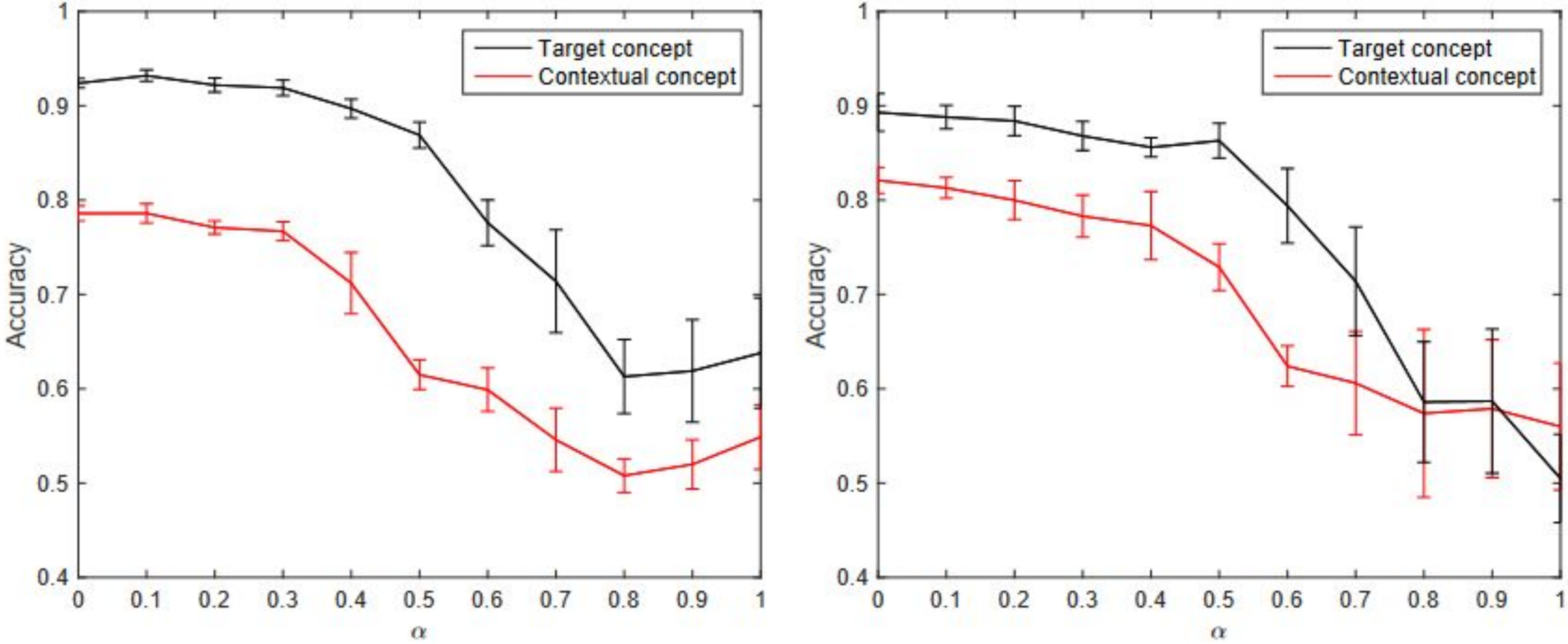
But think of machine justice or CV screening real cases … what is background there?



(a) Activation maps for the 'Killer whale' category

(b) Activation maps for the 'Jaguar' category

Fig. 4. Activation maps based on the strongest response of the shared feature representation. Examples selected are those with the least correlation between the activation maps for $\alpha = 0$ and $\alpha = 0.8$ as shown in the images.



(a) Performance of $G_y$ in the DANN trained for the target concept (animals)

(b) Performance of $G_p$ in the DANN trained for the contextual concept (backgrounds)

Fig. 3. Accuracy of the two independent classifiers in the DANN using the shared feature space on the test sets for different values of $\alpha$.

# About Transparency / Explainability

The right to an explanation

LOCAL vs GLOBAL explanations

Various projects under way …

# LIVING with intelligent machines...

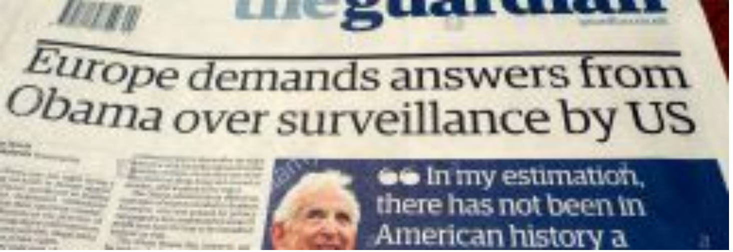
What started as a great idea to increase our autonomy, enable and empower users,  and by-pass some gate-keepers, has turned into a shift of power  to a new place, - to algorithms and companies

The shortcuts we took to AI are also the same choices that created many of the current problems

We are still deploying AI machines into our society and trusting them with personal decisions

We have MANY benefits from this technology, and we should not go back [all those listed at the start: whistleblowing, new companies, easier access to knowledge, ...]

BUT new laws, technologies  and ideas are needed before we can live safely with intelligent machines



theguardian

Europe demands answers from Obama over surveillance by US

In my estimation, there has not been in American history a

# … just there is no free lunch

An old story, about new technology…

## References

Image credit - Fair Use -

Images of News Headlines from:

NY Times, Washington Post, National Review, ProPublica, Wall Street Journal, The Guardian, USA Today, wired magazine, BBC, buzzfeed, science magazine, CNN

+ FBI website, Stanford Obituaries

Fair Use - images of products from Marketing material from: Cambridge Analytica, Admiral, amazon.com, apple.com, google.com, deepmind.com, nuzzel.com, trivago.com, axciom

Prism slide from: https://www.theguardian.com/world/interactive/2013/nov/01/prism-slides-nsa-document