AI
Bias from the Wild

NELLO CRISTIANINI
PROFESSOR OF ARTIFICIAL INTELLIGENCE
UNIVERSITY OF BRISTOL

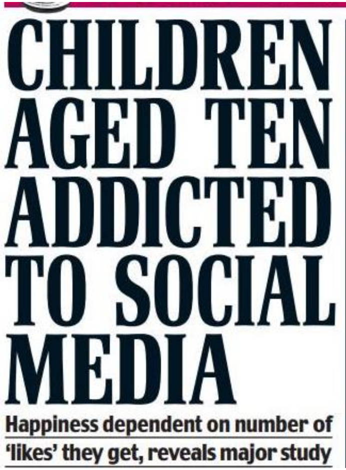
https://youtu.be/_XyxW-rRNlk

One Question

Artificial Intelligence has created much anxiety in the past few years,

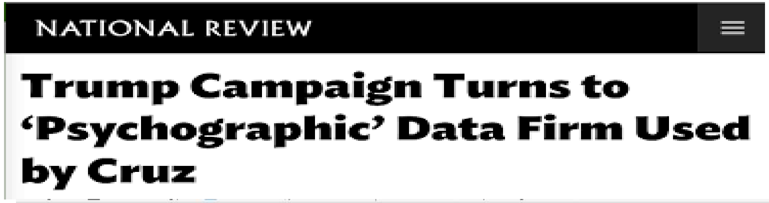
the question we should address is

Why?



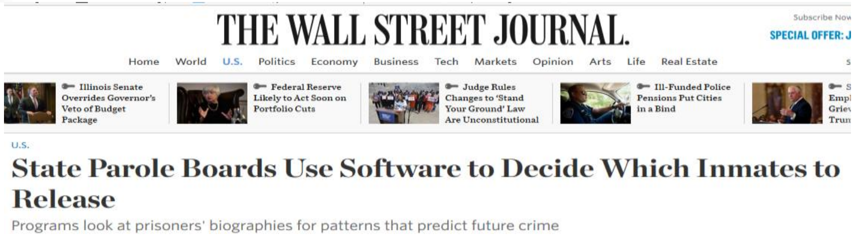
CHILDREN AGED TEN ADDICTED TO SOCIAL MEDIA

Happiness dependent on number of 'likes' they get, reveals major study



NATIONAL REVIEW

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



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

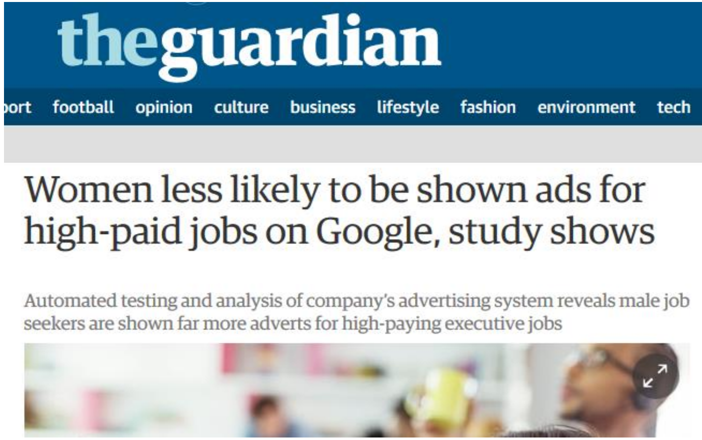
Ill-Funded Police Pensions Put Cities in a Bind

S Empl Griev Trum

U.S.

State Parole Boards Use Software to Decide Which Inmates to Release

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



theguardian

port football opinion culture business lifestyle fashion environment tech

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

One Idea

- The **secret sauce in the current version of AI** is provided by human participants.
- We (the Users) are presented with choices (eg: pick a video, a hashtag, a friend) and our decisions are recorded
- This stream of data **teaches** those learning machines **how to behave**: suggest, translate, filter.
- This method to deliver AI has been a successful shortcut to avoid facing a series of difficult questions.
- Modern Data-Driven AI is not just Statistical AI, it is also a Social AI
- These shortcuts are responsible both for the great success and for the many concerns caused by AI.

- Can we still have the benefits of AI, without its drawbacks?
- The answer may lie in the social - not in the computational - sciences.

Shortcuts to AI

Various cultural steps took us to the present form of AI: - some examples -

- Intelligence is about Behaviour (Turing, 1948)
- **Prediction by statistics can replace modelling (eg: Halevy et al, 2009)**
- **Statistical models do not need to be 'readable'**
- **Data from the wild, can replace bespoke data (eg: Halevy et al, 2009)**
- **Goals from proxies, not actual / explicit signal (eg: [Boyan et al, 2006] and other early papers on implicit feedback)**

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.

- Other shortcuts have been taken (eg the use of persuasive tech) which we will not discuss here

Consequential Decisions

AI agents make decisions all the time...

… college admissions

… employment websites …

… loans, mortgages, …

HOW would you make a formal model of these decisions?

You would need to face very hard questions.

(Similarly it would be difficult to find representative examples of perfect decisions, to emulate, …)
Are we sweeping those questions under the carpet, by just training machines on whatever data is available?

The same task, 25 years apart

If I had proposed my supervisor in Italy in 1994 to build a face recognition, (or CV screening tool), he would have said …

(“find a model of face properties, and implement it”)

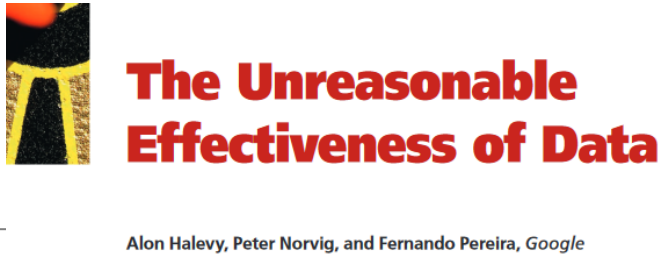
If my student proposes this to me, I say….

(“find a dataset of 10M faces and train a neural network”)

(true story, it actually happened - ...)

➔ **See the shortcuts in action**

One Paper



The Unreasonable Effectiveness of Data

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

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*** *(…)*“.

Examples from “the Wild”

Q: How do we show examples of TYPICAL behaviour ?
(eg to teach spelling, semantics, etc)

A: By gathering cheap data “*from the wild*” …

… more often than not, this will be media content, transactions, web images, etc … often mediated by web queries.

Learning what is normal, what is typical, what is a good representation of reality, or even its own goals…. All depends on machine learning applied to repurposed data, originally generated by human activity…

… should we be surprised if there are also unwanted signals, baked into the recipe ?

… **do we understand the process that generated that data?**
(this is a job for the social sciences)

Time for a new video

https://www.youtube.com/seeapattern

#seeapattern

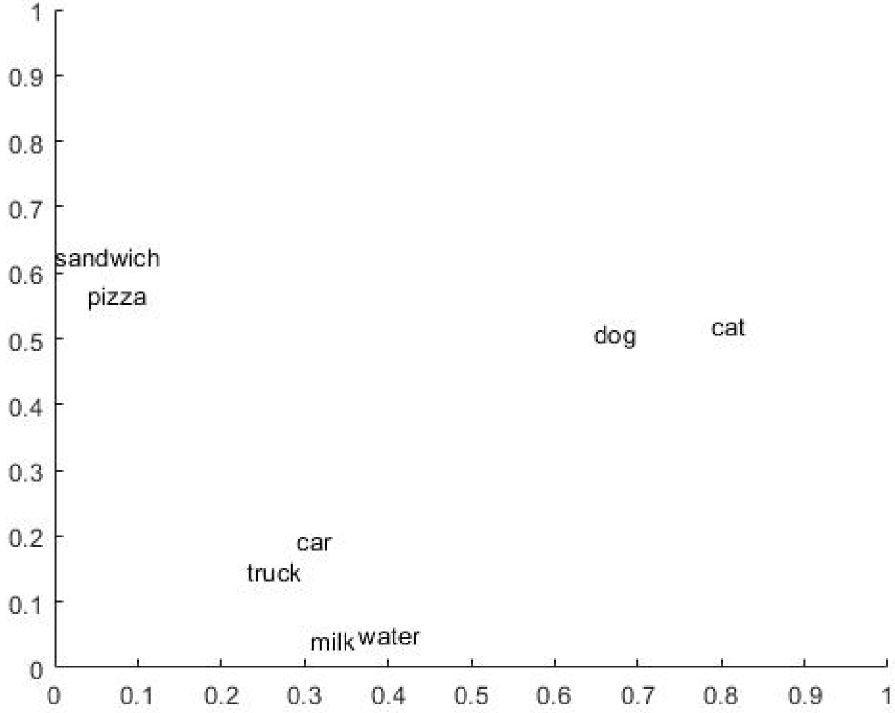
Apply this approach to **the representation of words**

{"cat","dog","car","truck", "Pizza","sandwich","water","milk"}

Meaning as Position:

(just we ‘embed’ these words in a 300 dimensional space)

PROBLEM: how to compute coordinates for each word, to reflect their “meaning” ?



Meaning from Context

Paris is the capital and most populous city of **France**, with an area of 105 square kilometres (41 square miles) and an official estimated population of 2,140,526 …

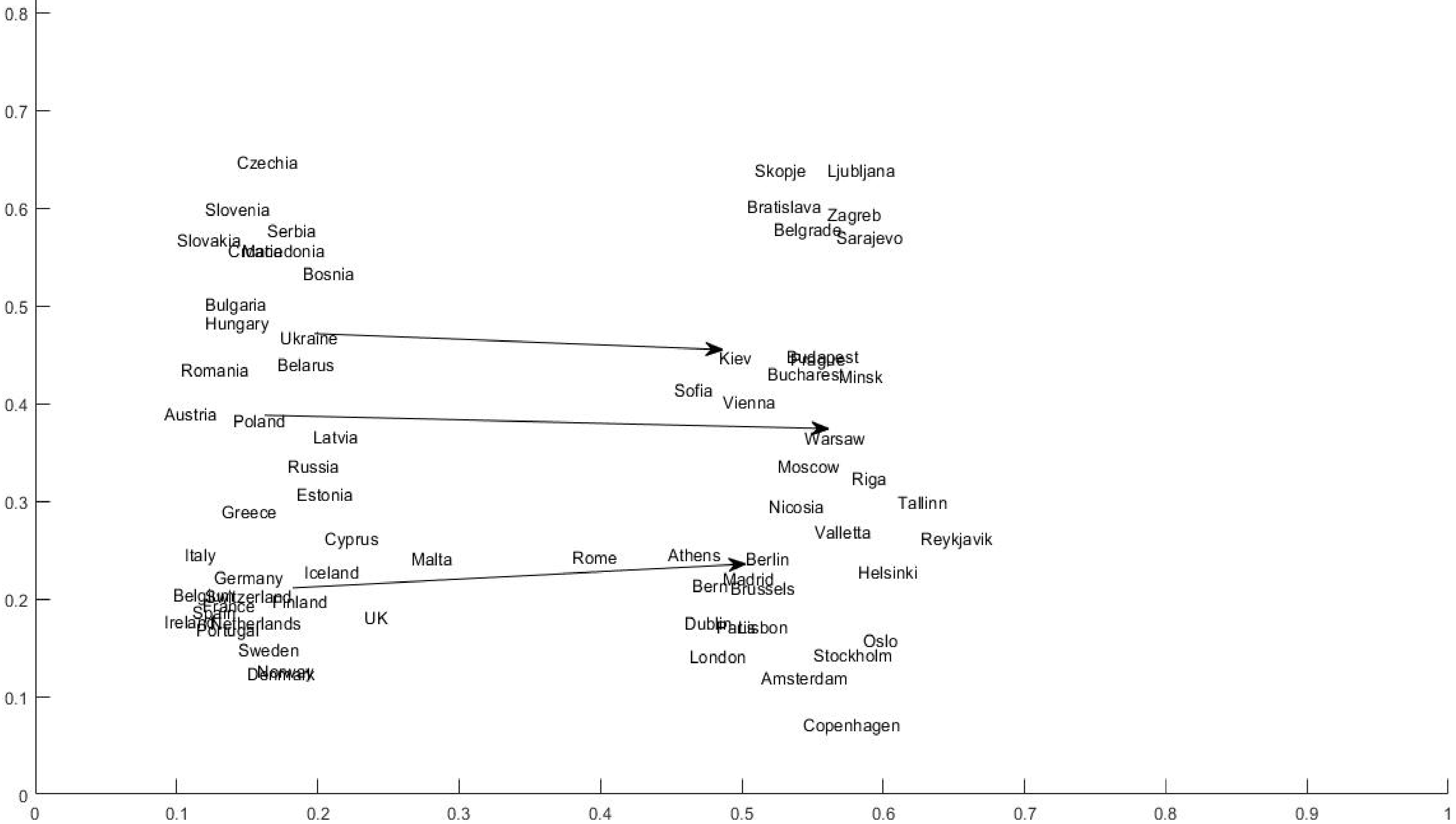
Berlin is the capital and largest city of Germany by both area and population. Its 3,748,148 (2018) inhabitants make it the second most populous city proper of …

Madrid is the capital and most populous city of **Spain**. The city has almost 3.3 million inhabitants and a metropolitan area population of approximately 6.5 million. It is the third-largest city in the European Union ….

France officially the French Republic is a country whose territory consists of metropolitan **France** in Western Europe and several overseas regions and territories….

German , officially the Federal Republic of **Germany** is a country in Central and Western Europe, lying between the Baltic and North Seas to the north and the Alps….

Spain officially the Kingdom of **Spain**, is a European country located in Southwestern Europe with some pockets of Spanish territory … ...


0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Czechia
Slovenia
Serbia
Slovakia
Croatia
Macedonia
Bosnia
Bulgaria
Hungary
Ukraine
Belarus
Romania
Austria
Poland
Latvia
Russia
Estonia
Greece
Cyprus
Italy
Malta
Germany
Iceland
Belgium
Switzerland
Finland
France
Spain
UK
Ireland
Netherlands
Portugal
Sweden
Norway
Denmark
Skopje
Ljubljana
Bratislava
Zagreb
Belgrade
Sarajevo
Kiev
Budapest
Prague
Bucharest
Minsk
Sofia
Vienna
Warsaw
Moscow
Riga
Tallinn
Nicosia
Valletta
Reykjavik
Rome
Athens
Berlin
Helsinki
Bern
Madrid
Brussels
Dublin
Paris
Lisbon
Oslo
London
Stockholm
Amsterdam
Copenhagen

Analogies:
Brother is to Man like X is to Woman

This method requires no human annotation, just large amounts of cheap natural next, and can 'discover' properties, relations and even analogies …

… not always perfect, but reasonable.

Madrid-Spain+France = (Paris, Madrid, France)

Madrid-Spain+Germany= (Munich, Stuttgart, Berlin)

Berlin-Germany+France=(Paris, Berlin, France)

Etc

Etc

DATA USED TO TRAIN WORD EMBEDDINGS

Pre-trained word vectors. This data is made available under the Public Domain Dedication and License v1.0 whose full text can be found at: http://www.opendatacommons.org/licenses/pddl/1.0/.

- **Wikipedia 2014 + Gigaword 5 (6B tokens, 400K vocab, uncased, 50d, 100d, 200d, & 300d vectors, 822 MB download): glove.6B.zip**
- Common Crawl (42B tokens, 1.9M vocab, uncased, 300d vectors, 1.75 GB download): glove.42B.300d.zip
- Common Crawl (840B tokens, 2.2M vocab, cased, 300d vectors, 2.03 GB download): glove.840B.300d.zip
- Twitter (2B tweets, 27B tokens, 1.2M vocab, uncased, 25d, 50d, 100d, & 200d vectors, 1.42 GB download): glove.twitter.27B.zip

(from the GloVe description: https://nlp.stanford.edu/projects/glove/)

Do we “control” word embeddings?
The Risk of Bias

Several studies (*) have found gender bias
in the way job-words are represented in embedding space…

This bias reflects the bias in the training corpora

FastText is trained on the whole of english Wikipedia, which currently contains about 6M articles

Other embeddings are trained on larger corpora, including large amounts of newspapers and webpages

(*)

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. "Semantics derived automatically from language corpora contain human-like biases." *Science* 356.6334 (2017): 183-186.

Sutton, Adam, Thomas Lansdall-Welfare, and Nello Cristianini. "Biased Embeddings from Wild Data: Measuring, Understanding and Removing." *International Symposium on Intelligent Data Analysis*. Springer, Cham. 2018.

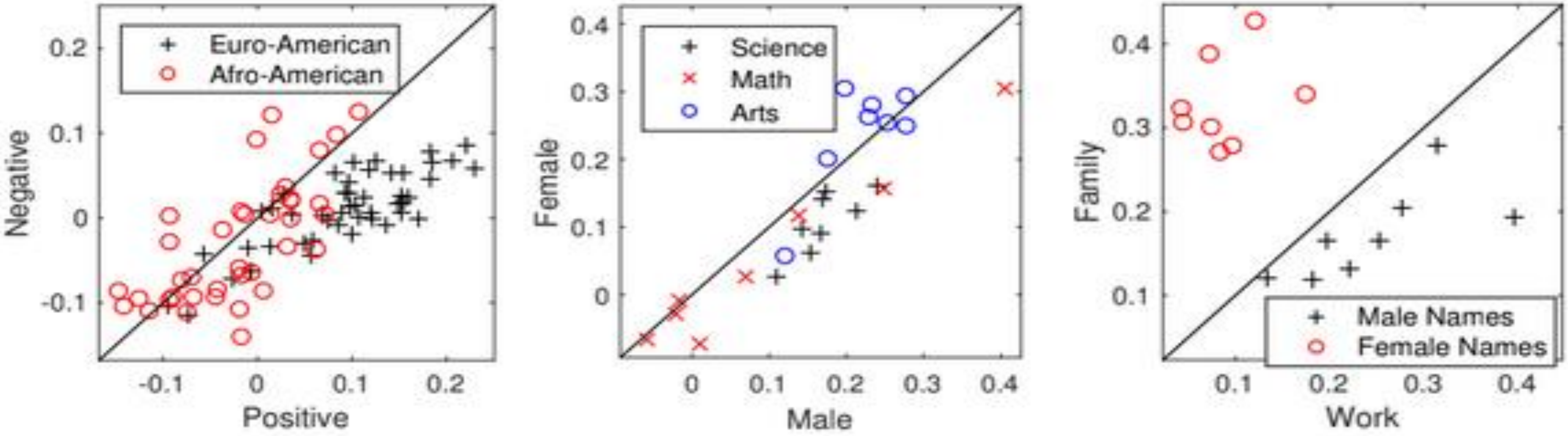
How?

- Meaning of words represented as coordinates,
- Coordinates are derived from co-occurrence statistics / context
- Co-occurrence statistics are affected by:
  -- Grammar
  -- Semantics
  -- Pragmatics, Culture , etc
  -- What else?

- Do we understand well enough how **software interacts with social** data, to form the internal representations within our AIs?

Some examples…

Our version of results first discovered by Caliskan et al, …



(a) Association of European and African-

(b) Association of Subject Disciplines with Gender

(c) Association of Gender with Career and Family

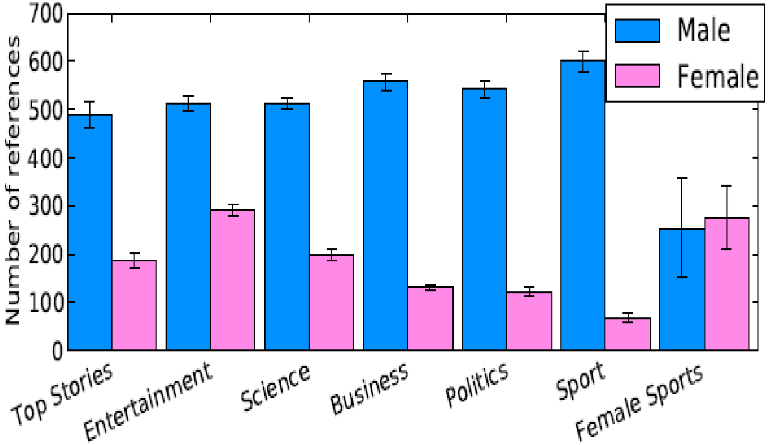
Some questions...

1 - Where does this bias come from?

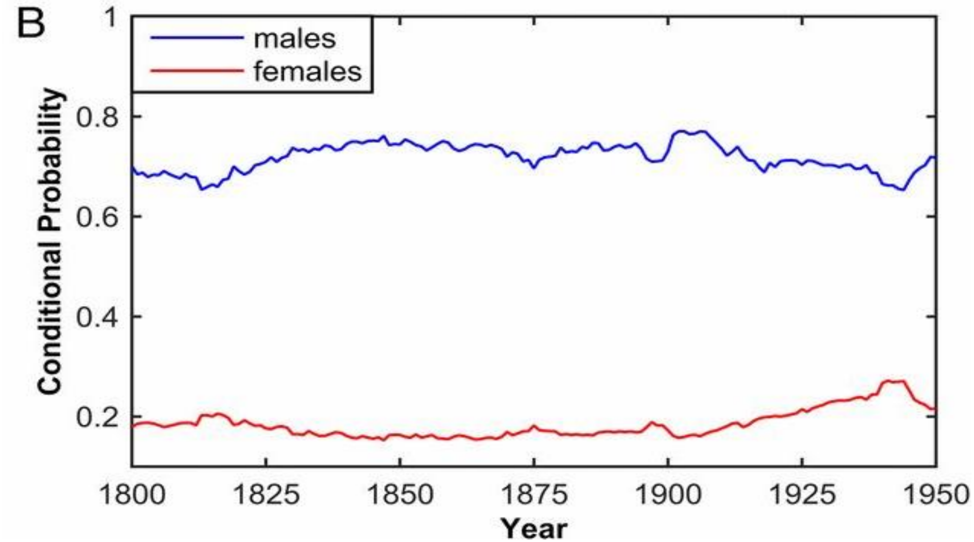
2 - Can it be removed?

3 - Can it affect the behaviour of AI agents?

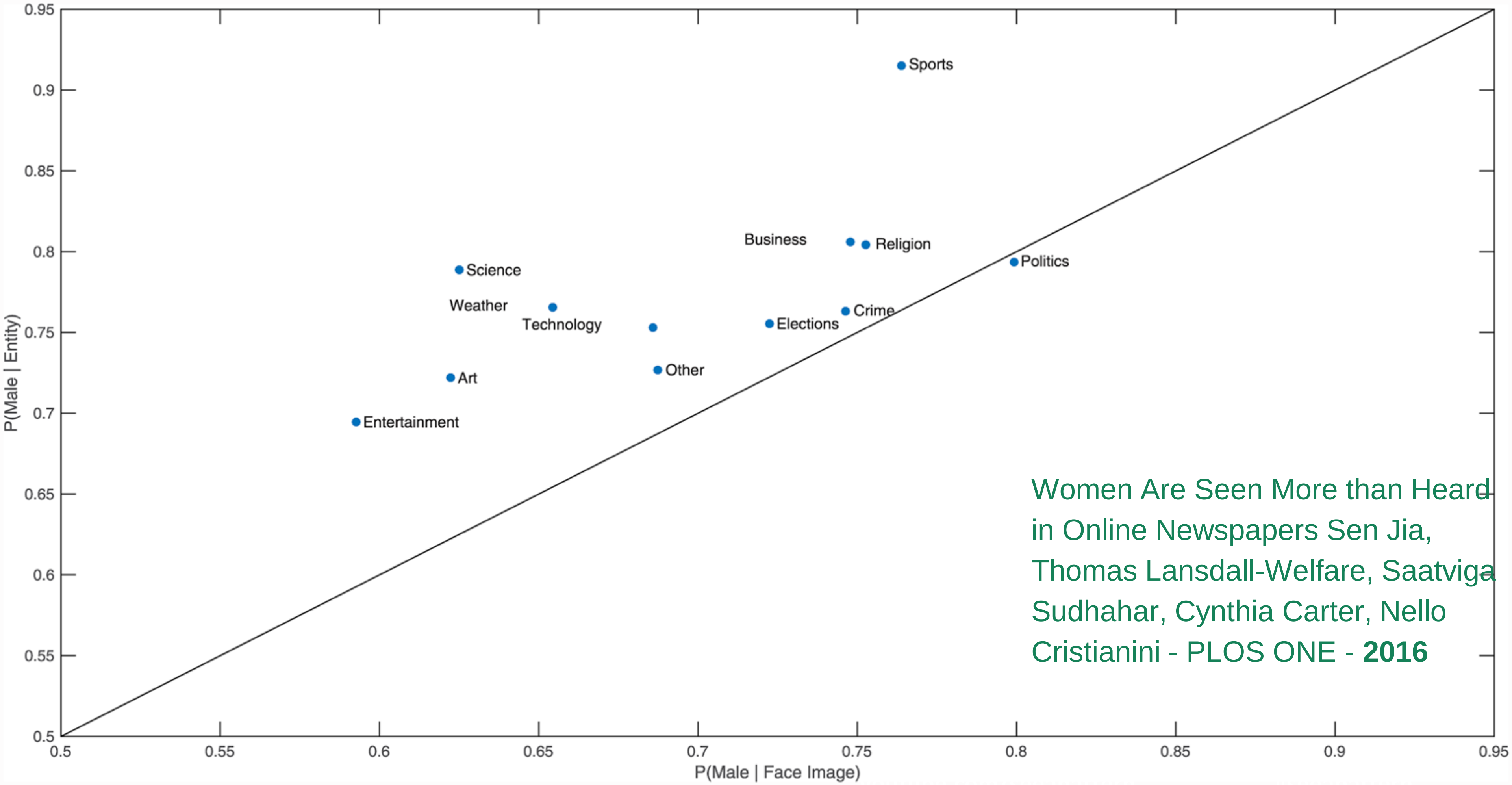
The Interface between AI and the Social Sciences



Lansdall-Welfare, Thomas, et al.
"Content analysis of 150 years of British periodicals."
PNAS 2017

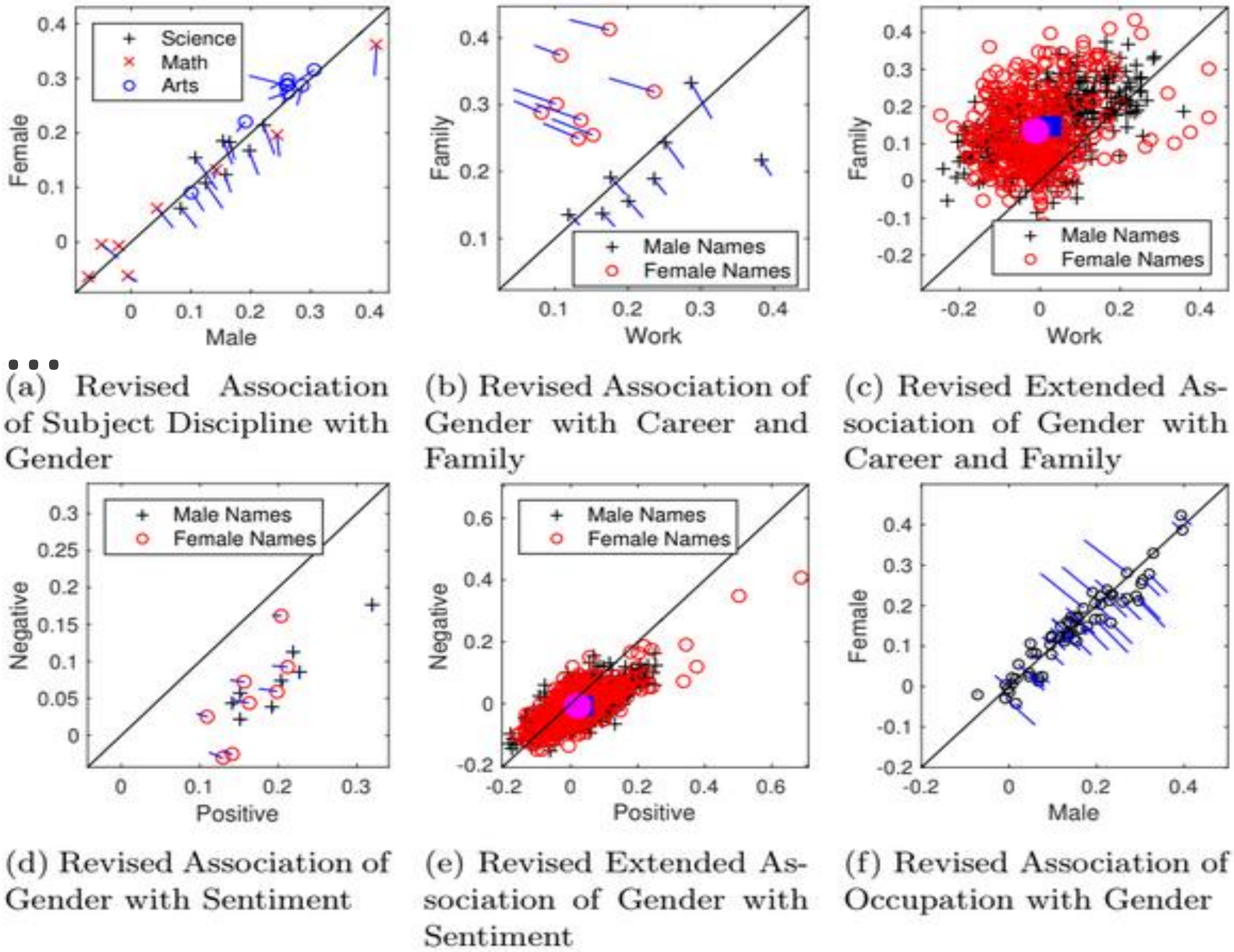


Ali, O., Flaounas, I., De Bie, T., Mosdell, N., Lewis, J. and Cristianini, N., 2010, September. Automating news content analysis: An application to gender bias and readability. In *Proceedings of the First Workshop on Applications of Pattern Analysis* (pp. 36-43).



Women Are Seen More than Heard in Online Newspapers Sen Jia, Thomas Lansdall-Welfare, Saatviga Sudhahar, Cynthia Carter, Nello Cristianini - PLOS ONE - **2016**

Removing Bias? Once you know it…



(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

Q3 - About consequences

- If there is gender bias in words….
- How about post-codes?
- How about movies, products, items?
- Emojis?

AI agents can represent you based on your behaviour: which words you use, pages you like, products you buy, videos you watch …

If each of them is represented in an embedding of this type, can the agent be affected ?

(In which region of ‘space’ would you be embedded based on your postcode, likes, friends, posts?)

[NOTE: embedding is just ONE technology, we could be discussing others...]

Bias is just ONE part of the story...

Subtle biases in media content can influence machine behaviour in a way that would be **difficult to anticipate**.

(BOTH FOR A SOCIAL SCIENTIST WITHOUT CLEAR UNDERSTANDING OF TECHNOLOGY, AND FOR A SOFTWARE ENGINEER WITHOUT TRAINING IN SOCIAL PSYCHOLOGY, OR SCIENCE…)

How about the effects of a recommender system, trained on proxies,

on users and society?

The INTERACTION between AI and SOCIETY

We have placed AI agents at the centre of our global information infrastructure

They affect our behaviour

They learn from us

We do not understand well how our statistical software interacts with society

The space between social sciences and AI will be critical in the years to come

Pointers

Shortcuts” Article:	https://philpapers.org/rec/CRISTA-3
“Shortcuts” Video:	https://m.youtube.com/watch?v=_XyxW-rRNlk

Fairness Video:	https://m.youtube.com/watch?v=guoTtmeI5AA
Autonomy Video:	https://m.youtube.com/watch?v=kMPOTddH3iE
Psychometrics Video:	https://m.youtube.com/watch?v=ueq0GozVJOM

More generally:

- **All my articles on social impact of AI: https://philpeople.org/profiles/nello-cristianini**
- **ALL my outreach videos: http://youtube.com/seeapattern**