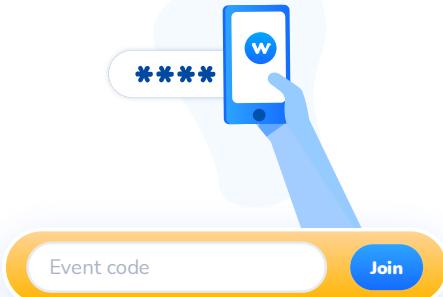


WARMUP

wooclap

Participate in an event



Choose a method to log in
Don't have an account? [Sign up](#)

@ Email

OR

OR

Log in with your institution
Only for institutions with a Custom Wooclap subscription

 Your institution ▾



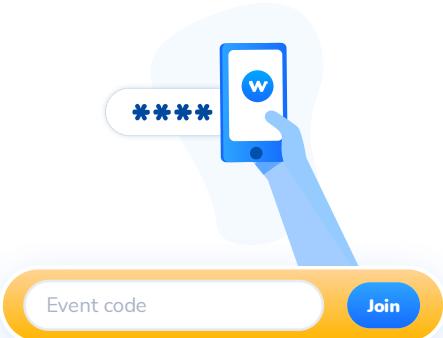
We use [cookies](#) to improve the general experience on the platform, provide users with chat support and deliver targeted ads on other websites.

[Accept all](#)

[Reject All](#)

[Customise](#)

Participate in an event



Event code

Join

Choose a method to log in

Don't have an account? [Sign up](#)

@ Email

OR



OR

Log in with your institution

Only for institutions with a Custom Wooclap subscription

 Your institution



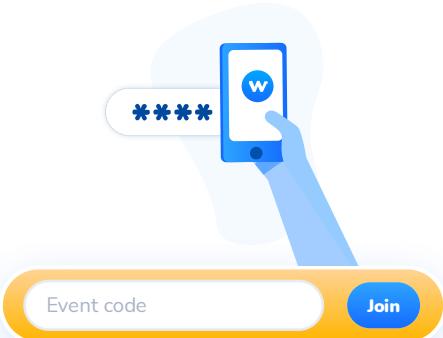
We use [cookies](#) to improve the general experience on the platform, provide users with chat support and deliver targeted ads on other websites.

Accept all

Reject All

Customise

Participate in an event



Event code

Join

Choose a method to log in

Don't have an account? [Sign up](#)

@ Email

OR



OR

Log in with your institution

Only for institutions with a Custom Wooclap subscription

 Your institution



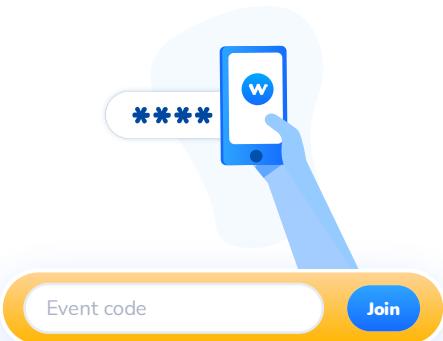
We use [cookies](#) to improve the general experience on the platform, provide users with chat support and deliver targeted ads on other websites.

Accept all

Reject All

Customise

Participate in an event



Event code

Join

Choose a method to log in

Don't have an account? [Sign up](#)

@ Email

OR



OR

Log in with your institution

Only for institutions with a Custom Wooclap subscription

 Your institution



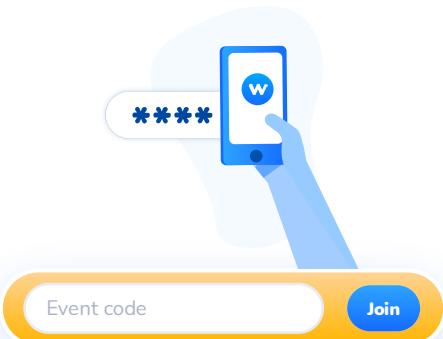
We use [cookies](#) to improve the general experience on the platform, provide users with chat support and deliver targeted ads on other websites.

Accept all

Reject All

Customise

Participate in an event



Event code

Join

Choose a method to log in

Don't have an account? [Sign up](#)

@ Email

OR



OR

Log in with your institution

Only for institutions with a Custom Wooclap subscription

 Your institution



We use [cookies](#) to improve the general experience on the platform, provide users with chat support and deliver targeted ads on other websites.

Accept all

Reject All

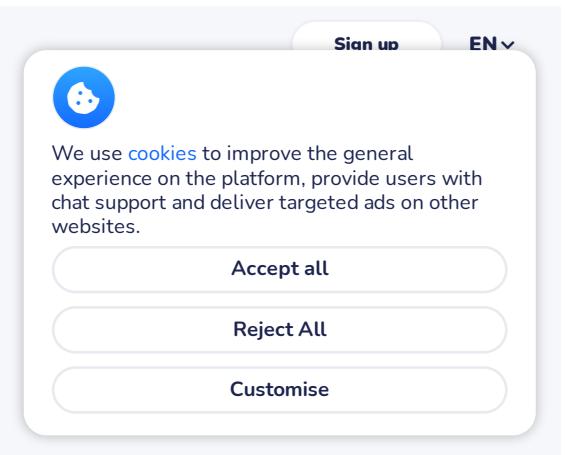
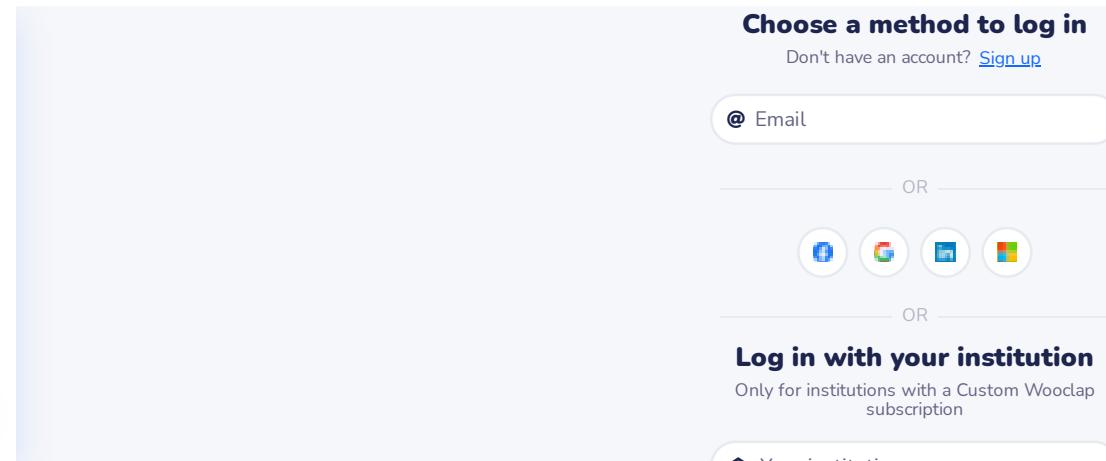
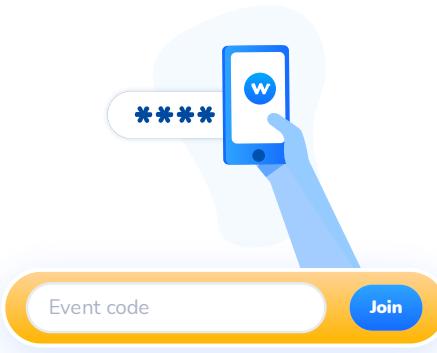
Customise

AI BIASES VS HUMAN BIASES

PABLO WINANT, ESCP BUSINESS SCHOOL

A SMALL QUESTION

wooclap **Participate in an event**



- Are you ready to be driven by an AI-driven car, 5 years from now?
- Info about car accidents (today)
 - AI: 9 crashes per million mile
 - human: 4 crashes per million mile
 - but almost no major injury in AI driven cars
- AIs are easy to fool
 - incorrect reading of traffic signs with small modifications

AIS WILL TAKE MORE AND MORE DECISIONS

- AIs will take more and more decisions
 - decide what you'll watch on Netflix
 - drive your car
 - select the recruits you will hire
 - decide whether you should be receiving treatment from the nearby hospital
 - invest your personal finances
 - decide optimal monetary policy of the central bank
- But there will always be a human overseeing these AI decisions?
- ...right?

WHAT IS A DECISION

- Several seemingly different cases:
 - recommendation
 - decision with immediate consequences
 - a part of a decision process
- These cases are not so clearly separable
- Precise agency is not important here
- We'll call of these "decisions"
 - (alternatives: "predictions"/"choices"/...)

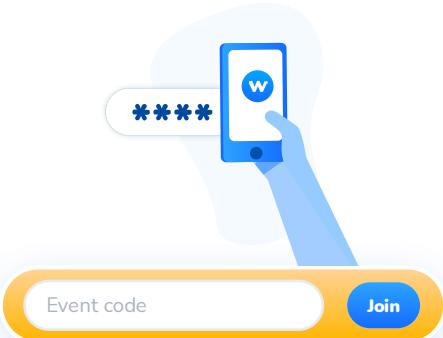
Cassie Kozyrkov, chief decision scientist from Google:

We define the word “decision” to mean any selection between options by any entity

DECISION INTELLIGENCE

- Who studies AI decisions?
- A new emerging field: "Decision Intelligence"
- Defines "decisions" as
 - a choice of an "output" from a set of "input"
 - choice is irreversible
- Relates data-science with different fields

Participate in an event



Event code

Join

Choose a method to log in

Don't have an account? [Sign up](#)

@ Email

OR



OR

Log in with your institution

Only for institutions with a Custom Wooclap subscription

 Your institution



We use [cookies](#) to improve the general experience on the platform, provide users with chat support and deliver targeted ads on other websites.

Accept all

Reject All

Customise

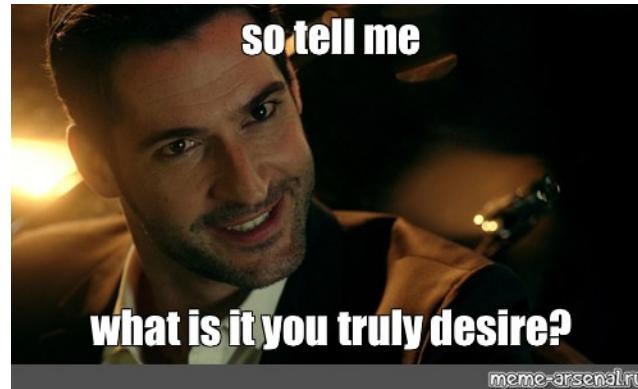
Example of questions: (from Cassie Kozyrkov, chief decision scientist from Google)

The decision sciences concern themselves with questions like:

- “How should you set up decision criteria and design your metrics?” (All)
- “Is your chosen metric incentive-compatible?” (Economics)
- “What quality should you make this decision at and how much should you pay for perfect information?” (Decision analysis)
- “How do emotions, heuristics, and biases play into decision-making?” (Psychology)
- “How do biological factors like cortisol levels affect decision-making?” (Neuroeconomics)
- “How does changing the presentation of information influence choice behavior?” (Behavioral Economics)
- “How do you optimize your outcomes when making decisions in a group context?” (Experimental Game Theory)
- “How do you balance numerous constraints and multistage objectives in designing the decision context?” (Design)
- “Who will experience the consequences of the decision and how will various groups perceive that experience?” (UX Research)
- “Is the decision objective ethical?” (Philosophy)

THE SPECIFICITIES OF AN ECONOMIC APPROACH

- Consider different ways to analyse AI behaviour from an economic perspective (*people's decisions*)
 - deviations from rationality
 - specification of a precise objective



THREE KINDS OF BIAS

- predictive bias
 - ... we know what want but are doing it wrong
- preference bias
 - ...we're wrong about the ultimate objective
- behavioural bias (next week)
 - ...we're doing it wrong

PREDICTION BIAS

DEFINITION OF STATISTICAL BIAS

Prediction Bias: systematic error made by an algorithm producing a prediction

- Here, *systematic* must be understood as *in average* or *in expectation*

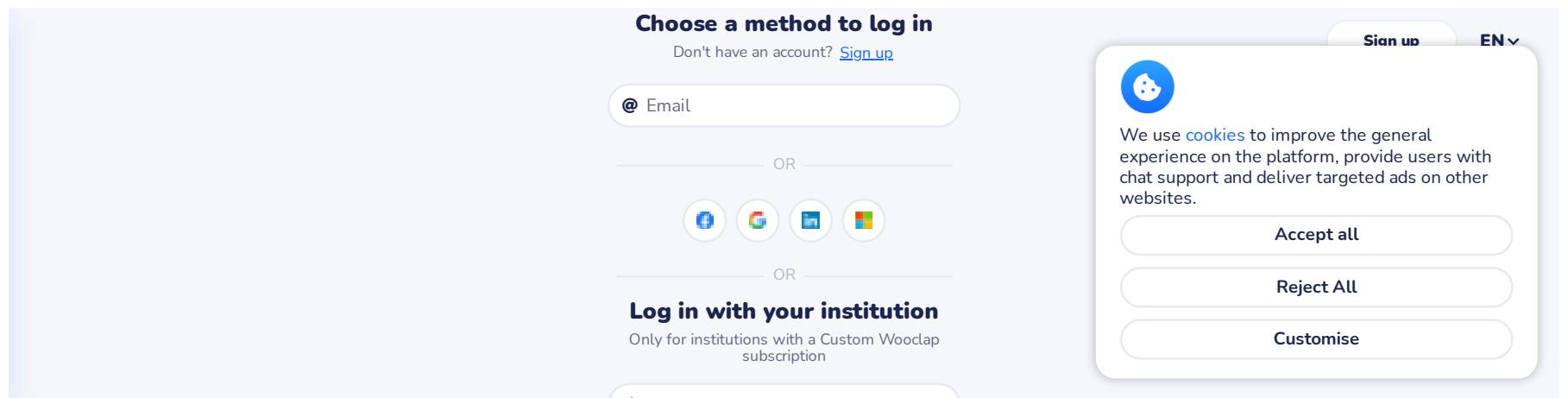
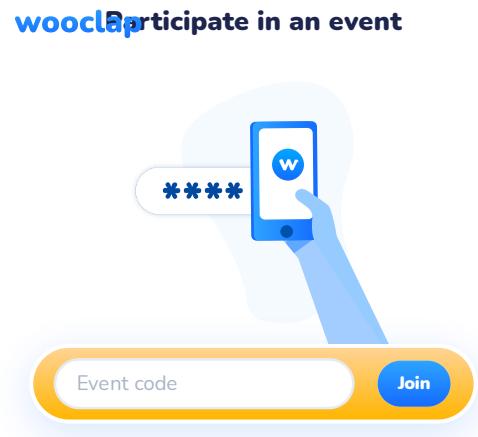
SOURCES OF PREDICTION BIASES

- Problems with the data (*data-driven*)
 - **selection bias**
 - **attrition bias**
 - ...
- Problems with the model (*algorithmic bias*)
 - **ommitted variable bias**
 - ...
- Other sources (essentially *human bias*)
 - **funding bias**
 - **social prejudice**
 - **human limitation**
 - ...

IMAGE LABELLING

An AI or you must choose labels to best describe this image:





The image shows the Wooclap login screen. At the top, it says "Choose a method to log in". Below that is a "Sign up" link. A "Email" input field is followed by an "OR" separator. Below "OR" are icons for LinkedIn, Google, Microsoft, and another social media platform. Another "OR" separator follows. Below that is a "Log in with your institution" section, which is noted as being for institutions with a Custom Wooclap subscription. At the bottom right is a cookie consent banner with a blue circular icon. It states: "We use cookies to improve the general experience on the platform, provide users with chat support and deliver targeted ads on other websites." It includes three buttons: "Accept all", "Reject All", and "Customise".

AN EXAMPLE OF A SELECTION BIAS



Obviously, the way the AI (or you) makes category, depends on the dataset it has been exposed to.

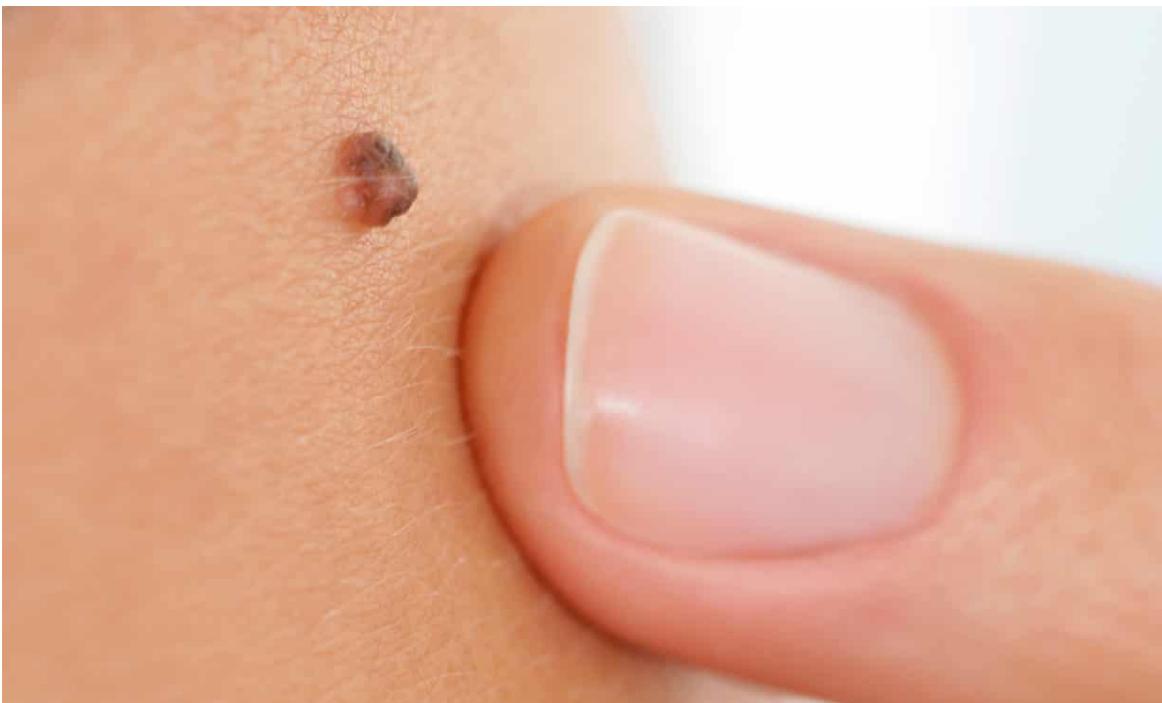
- Experience/learning produces a prototype of a watermelon (i.e., an object representative of its class)

HOW DO WE MEASURE IT ?

- Sometimes bias is easy to measure with
 - precise criterium (e.g. no discrimination)
 - precise measure (e.g obvious distribution discrepancies)
- But in general it requires:
 - an experiment
 - some econometric work
- Often, biases are easier to assert for AIs than humans
 - their training occurs in a controlled environment

DO YOU WANT TO BE TREATED BY AN AI?

[Nature, 25/01/2017](#): Dermatologist-level classification of skin cancer with deep neural networks



- analyze skin images to recognize malignant melanoma
- as good as human dermatologists
- more cost-effective (can work on a smartphone)

DO YOU (REALLY) WANT TO BE TREATED BY AN AI?

[The Lancet, 2022](#): *Characteristics of publicly available skin cancer image datasets: a systematic review*

- They review 21 open access databases, with skin lesion images
 - 106 950 images
 - Of the two datasets containing data on ethnicity (1585 images in total), 45, 47 no images were from individuals with an African, Afro-Caribbean, or South Asian background
 - Coupled with the geographical origins of datasets, there was massive under-representation of skin lesion images from darker skinned populations.
- Conclusion?

OR DO YOU PREFER TO BE TREATED BY A HU(MAN) ? (1)

Health Services As Credence Goods: A Field Experiment

(Gootschalk, Mimra, Weibel)

- The same "test patient" was sent to 180 dentists who offered treatment recommendation and cost estimate.
- Test patient did not need treatment (caries lesions limited to enamel).
- 28% of practitioners made a wrong treatment recommendation! 😱
- What were the determinants of the bias?
 - Social Economic Status (-)
 - Lower Waiting Time (+)

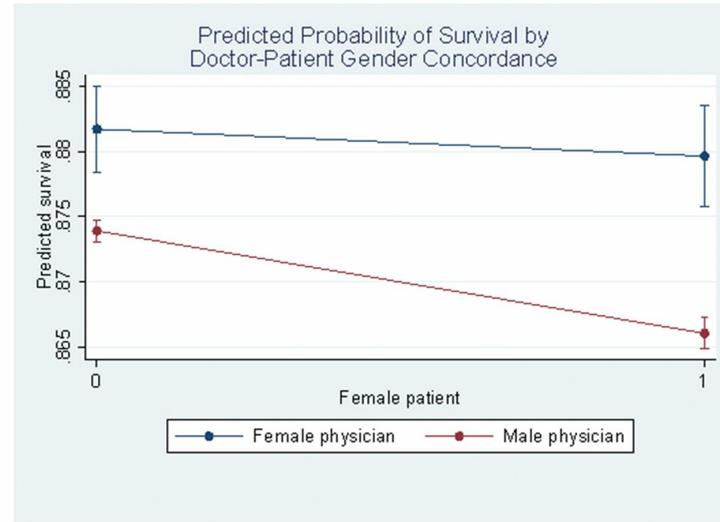


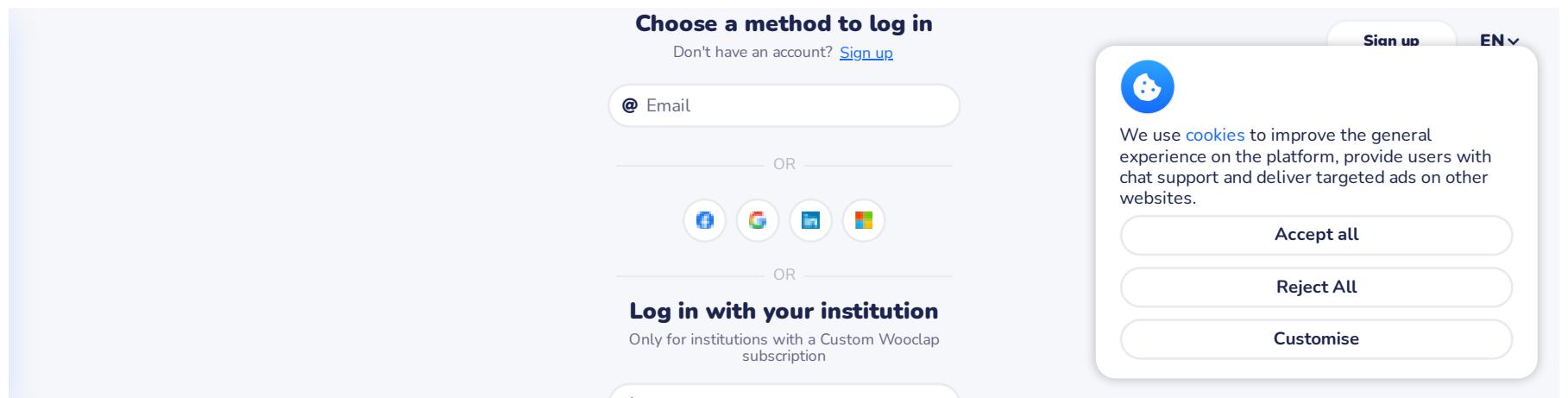
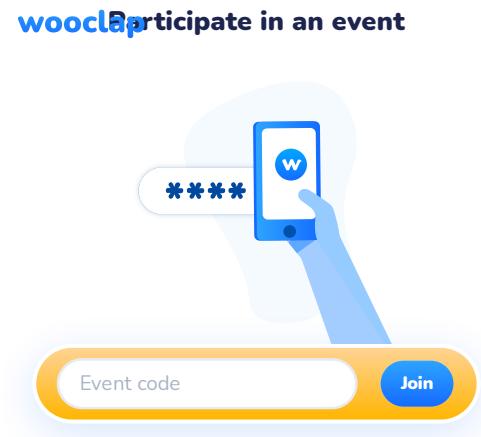
OR DO YOU PREFER TO BE TREATED BY A HU(MAN) ? (2)

Perceived Risk of Heart Attack: A Function of Gender?
2004, (Leanne L Lefler)

Patient-physician gender concordance and increased mortality among female heart attack patients
(Greenwood, Carnahan, Huang)

- mortality rate for women in the year immediately after suffering a heart attack was 38%, compared to 25% for men
 - woman delay assistance seeking (it's a men problem)?
- higher probability of survival when same-sex doctor
 - driven by treatment from male doctors (the majority of cardiologists)





The image shows the Wooclap login screen. At the top, it says "Choose a method to log in". Below that is a "Sign up" link. A "Email" input field is followed by an "OR" separator. Below "OR" are icons for LinkedIn, Google, Microsoft, and another social media platform. Another "OR" separator follows. Below that is a "Log in with your institution" section, which is noted as being for institutions with a Custom Wooclap subscription. At the bottom right is a cookie consent banner with a blue circular icon. It states: "We use cookies to improve the general experience on the platform, provide users with chat support and deliver targeted ads on other websites." It includes three buttons: "Accept all", "Reject All", and "Customise".

MARKET BIAS

- Job Market
 - *Job discrimination*: the decision to hire someone at a given salary should not depend on his/her gender, appearance, social origin, age, ethnicity, ...
 - *Wage gap*: also the wage gap between people with the same overall productivity should be zero, no matter their gender, appearance, ...
- Big problems:
 - how do you measure "same overall productivity"?
 - wrong measurement leads to **ommitted variable bias**
 - how do you find two people with different characteristics and exactly same productivity?
 - in general many characteristics are linked with productivity (for instance, name is correlated with IQ)

A NOT SO-FICTIVE EXAMPLE OF A FAILED ANTI-DISCRIMINATION POLICY



Bob from Texas

FAMOUS EXAMPLE: AMAZON

[Reuters](#) 11/10/2018: Amazon scraps secret AI recruiting tool that showed bias against women



- What happened?
 - Amazon started to train (use?) internally a ML algo to preselect CVs and counteract human biases
 - Algorithm started to discriminate against woman
 - Sentences containing strings like "women's" were discriminated against (like "champion of women's chess cup")

ANOTHER HUMAN BIAS



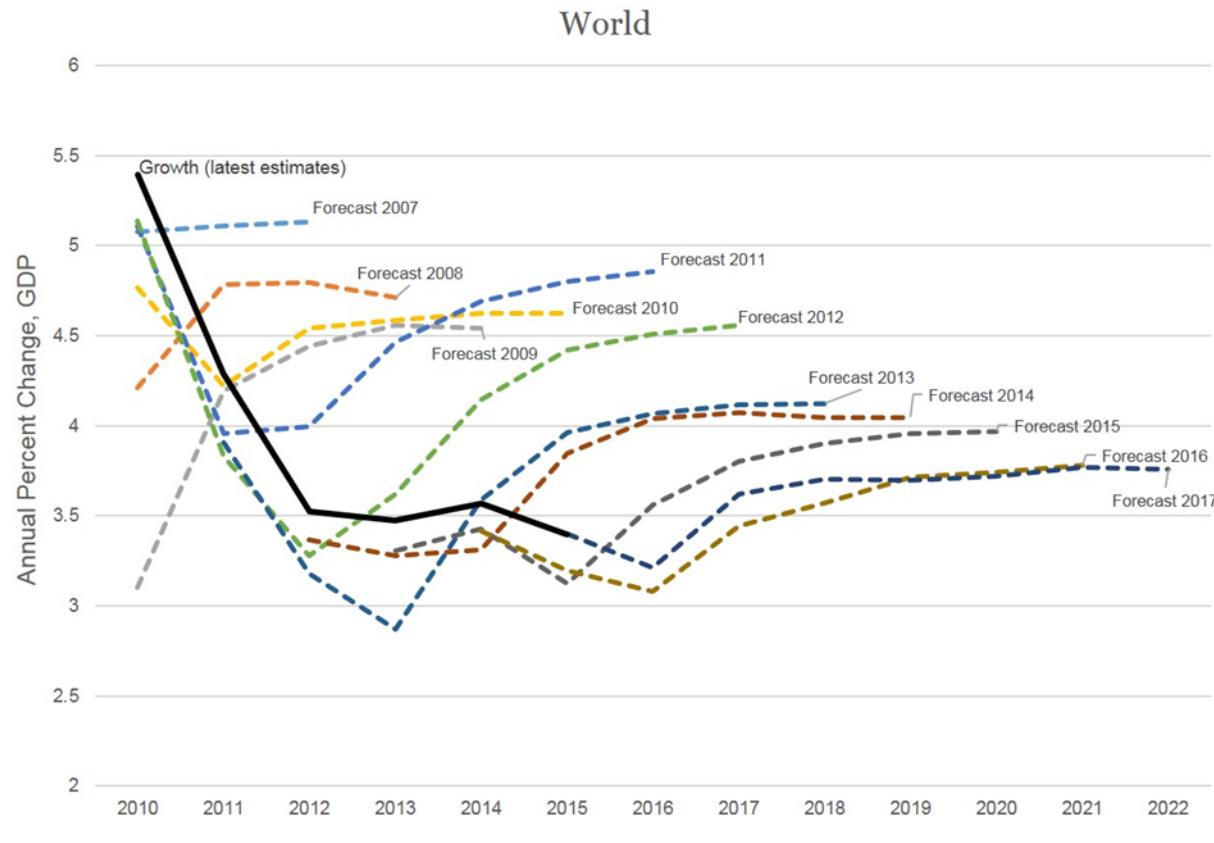
- PNAS 2017, *Extraneous factors in judicial decisions*, Danziger, Levav and Avnaim-Pesso
 - analyse parole decisions by boards presided by jewish isreali judges
 - probability of "parole" falls between the two snack breaks! 😱
 - find a strong "hungry judge effect"

CONCLUSIONS

- AI and humans *both* suffer from **selection bias**
- AI can reproduce human biases
 - in the way algorithm is designed
 - if it imitates humans or if its objective incorporates human bias, conscious or not
- AI's don't have all human biases
 - no hungry judge effect
 - no funding cost (or do they?)

ANOTHER EXAMPLE: WEO FORECAST

Here is the history of IMF forecasts (form WEO 2017)



Is it biased?

PREFERENCE MISSPECIFICATION

WHAT IS THE RIGHT WAY TO DESCRIBE ECONOMIC BEHAVIOUR?

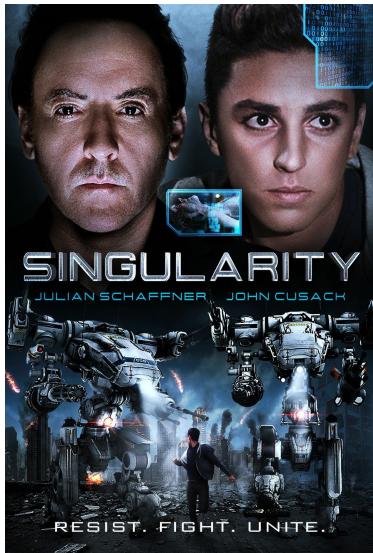
- In economics, we derive agent's behaviour from their ultimate objective
 - maximize profits
 - maximize consumption, leisure
 - something else
- This is very close to the implementation of AI now:
 - ML: miniminize empirical risk (sum of square residuals), maximize the fit
 - AI: robots are explicitly told what to do (not how)
- Biases should be precisely defined w.r.t. a well specified goal

EXAMPLE: BREXIT



- Was the collective decision of leaving the UK biased, based on available evidence?
- Here, the objective might not be well specified. There are unsaid, unconscious, objectives

EXAMPLES: AI OBJECTIVE MISSPECIFICATION

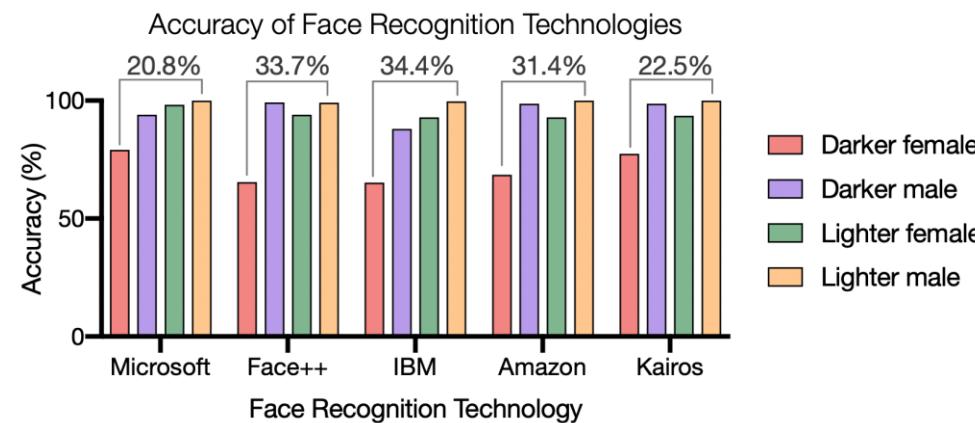


- AI with the wrong objective is evil
- Explored a lot in science-fiction
 - Asimov's law of robotics
 - the smiling man, ...
- Many (all?) "mistakes" from AI come from preference misspecification

EVOLUTIONARY BIAS

- It is sometimes optimal to keep a biased decision process in certain situations
- When taking bias decisions provides a survival advantage, it is called an **evolutionary bias**
 - treat unknown species as "hostile"
- Rationals
 - Limit processing cost
 - Provide informational value, i.e. help to learn faster
 - ...

AN EXAMPLE OF "TRIMMING"



- AI algorithm have become very good at recognizing and distinguishing faces...
 - ... mostly white men
 - selection bias again
- Adults have the same biases: they distinguish better faces from their own reference group
- Strikingly 6 month old babies don't: they recognize all faces (Netflix; "babies")

AN EXAMPLE OF LEARNING EXTERNALITY

- Why do newer movies have better ratings than older ones on movie databases (like Allocine)
- And why are website not doing anything about it?
- New movies are intentionnaly overrated
 - to push consumers towards "exploring"
 - to produce more information
 - and improve the rating of new movies
- It can be interpreted as a learning externality

PREFERENCES VS UTILITY

- Another issue is that *humans are not one-dimensional maximizers*
- Theories of **preferences** are larger than utility maximization
 - Among choices \mathcal{X} , we say that x is preferred to y if $x \geq y$
- Preferences can be more general than utility maximization
 - ideally transitive if $x \geq y$ and $y \geq z$ then $x \geq z$
 - but there isn't necessarily a total order (complete ranking) $x_1 \geq \dots \geq x_n$
 - even if there is there is no notion about "how much" x is preferred to y
- Generalized Preferences arise naturally from
 - real-world individuals
 - multi-objective agents
 - collective choices (cf Arrow Theorem)

MULTI-OBJECTIVES

- Ideal AIs should be multi-objectives:
 - have sensible default for out of sample **situations**
 - produce more intelligent behaviour
 - mitigate wrong objectives given by humans
- The problem is when AIs are trained to follow multiple objectives (which they need if they need a notion of context) their bias become harder to measure/explain

EXAMPLE: PARCOURSUP, A MULTI-OBJECTIVE RANKING ALGORITHM

- parcoursup match universities wishes and students wishes
 - it aggregate individual's preferences
 - while respecting current laws
- it is a variant of a stable marriage problem
- how do you formulate the optimum?
 - impossible to satisfy everybody
- implementation details makes random decisions
 - in order to avoid bias!
 - and satisfy local regulations
- has created a lot of discontentment

CONCLUSIONS

- The concept of predictive bias is contingent to the right, scalar (i.e. a number), objective specification
- Formulating the wrong the objective leads to a prediction bias
- Objectives are inherently hard to formulate exhaustively
- The presence of several objectives complicates the pictures
 - for humans
 - and AIs

BEHAVIOURAL BIASES

- ... Next time:
 - how do we know when humans act non-rationally?
 - Can we then establish some patterns in their behaviours?

FINAL WORD

It's good to follow your own bias as long as it is climbing it.

Andre Gide