AI BIASES VS HUMAN BIASES

PABLO WINANT, ESCP BUSINESS SCHOOL

A SMALL QUESTION

- 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 not major injury in AI driven cars
- AIs are easy to fool
  - incorrect reading of traffic signs with small modifications
  - see nature

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"/...)

DECISION INTELLIGENCE

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

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)

TODAY

We'll consider different ways to analyse AI behaviour from an economic perspective. In particular, we'll draw parallels, between AI decisions and human decisions

- biases from a quantitative/statistical approach
- the problem of preference misspecification
- behavioural mistakes (not today)
- homework, talk about your classwork

QUANTITATIVE BIAS

DEFINITION OF STATISTICAL BIAS

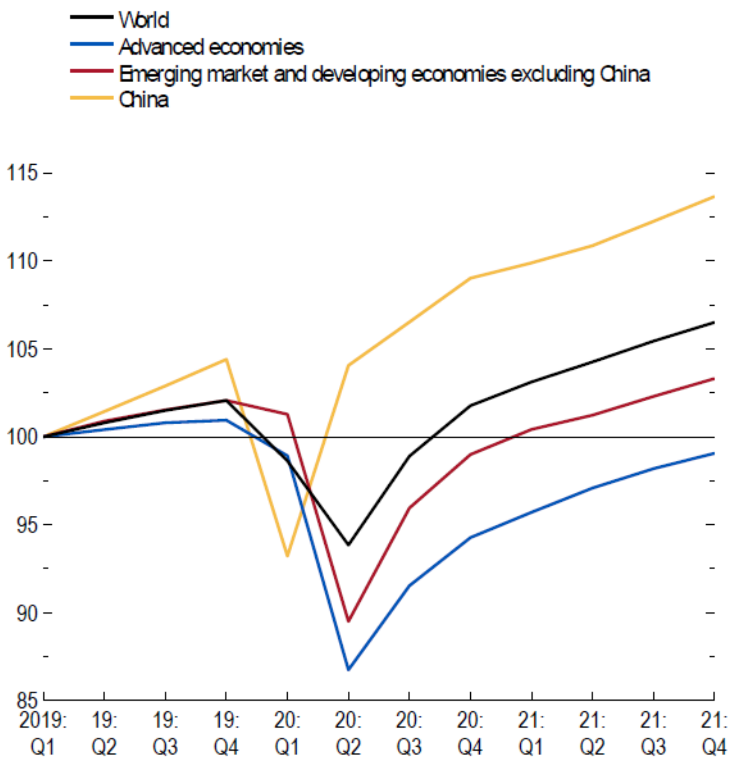
Bias: systematic error made by a a statistical algorithm producing a prediction

Here, *systematic* means, *in average*. (more precisely, in expectation w.r.t to all the sources of randomness).

EXAMPLE: WEO FORECAST

Here is the forecast from the latest World Economic Outlook (IMF)

Figure 1. Quarterly World GDP
(2019:Q1 = 100)



Source: IMF staff estimates.

Is it biased?

SOURCES OF BIASES

- Problems with the data (*data-driven*)
  - **Selection bias** / attrition biases
  - ...
- 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 needs to label best describe the following image:

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

- fruit

- Experience/learning produces a **prototype** of a

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

EXAMPLE OF 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*: conversely, 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"?
  - if you do, how do you find two people with different characteristics and exactly same productivity?
    - given that in general charecteristics and productivity are linked (for instance, name is correlated with IQ)
- One possibility: look at submitted CVs

AN EXAMPLE OF A FAILED ANTI-DISCRIMINATION POLICY

- Initial situation: Bob recruits new hires himself
  - he's got prejudice against: single women, obese men, non christian workers, ...
  - he drops unwanted CVs based on:
    - photographs
    - names
- New situation: Bob uses machine learning to select candidates who get an interview
  - task of ML: reject 95% of candidates
  - objective: maximize probability of that selected candidates get the job after their interview
  - diversity requirement: don't use name, gender and photo
- Result: after a few iterations, algorithm selects only young white candidates with christian names

What happened?

Algorithm has learned bias of user, and made it more efficient.

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")

EXAMPLE: 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)

EXAMPLE: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 (+)

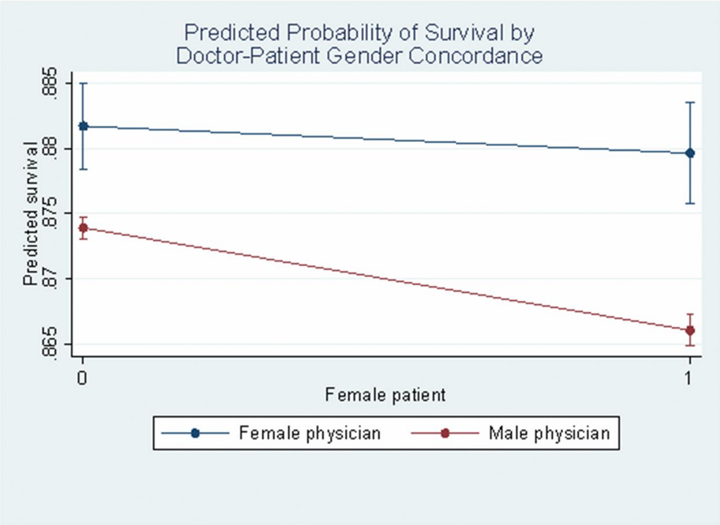
EXAMPLE:OR DO YOU PREFER TO BE TREATED BY A HU(MAN) ? (2)

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

- 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)?

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

- higher probability of survival when same-sex doctor
- driven by treatment from male doctors (the majority of cardiologists)



CONCLUSIONS

- AI can reproduce human biases
  - in the way algorithm is designed
  - if it immitates 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?)
- Humans also suffer from many of the same biases as machines
- Machines have some advantages
  - efficiency

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 explicitely told what to do (not how)
- Biases are 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

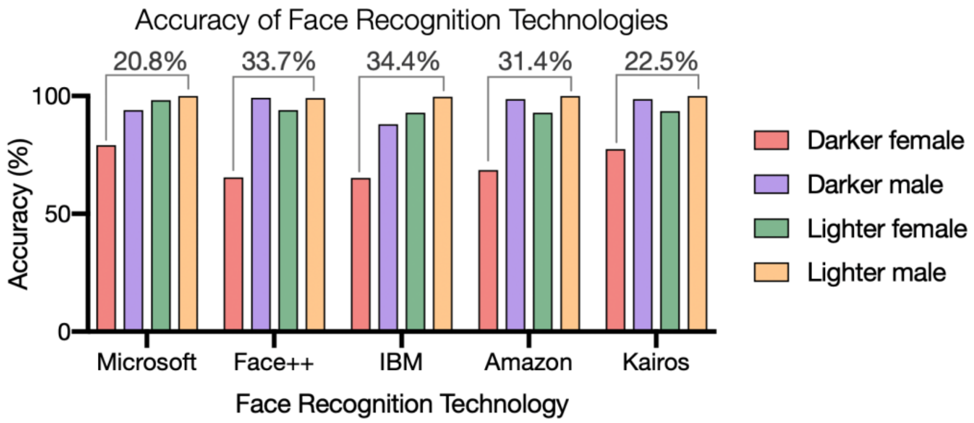
EXAMPLES: AI OBJECTIVE MISSPECIFICATION

- AI objective misspecification
  - famous scifi examples: Asimov's robots, the smiling man, ...
  - example: intertemporal consumption maximization

EVOLUTIONARY BIAS

- Under some circumstances, taking bias decisions can provide a survival advantage
  - treat unknown species as "hostile"
- 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 or
  - 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) $x_1 \geq \cdots \geq x_n$
  - 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

- We want multi-objectives:
  - have sensible default for out of sample situations
  - mitigate wrong objectives given by humans
- The problem is when AIs are follow multiple objectives (which they need if they need a notion of context) their bias becomes harder to measure

EXAMPLE: PARCOURSUP, A RANKING ALGORITHM

- parcoursup match universities wishes and students wishes
  - 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 bias is contingent to the right, scalar, objective specification
- That one is sometimes hard to formulate completely
- The presence of several objectives complicates the pictures
  - for humans
  - and AIs

YOUR PROJECT

COURSEWORK PROPOSITION

- a big advantage of AIs is that they can be tested easily
- if we had access to a general purpose AI, we could design experiments in order to test:
  - what are its revealed preferences (consistent, risk averse, irrational)
  - what biases it has
  - whether it exibhits similar behavioural biases than humans
- turns out we have such an AI: GPT-3
- your task:
  - assemble a 5 members max team
  - brainstorm about a creative way to study GPT-3 behaviour
    - choose any angle you want
  - think about an experimental protocol
  - carry it on if you can
  - present it as if it was a research project

FINAL WORD

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

Andre Gide