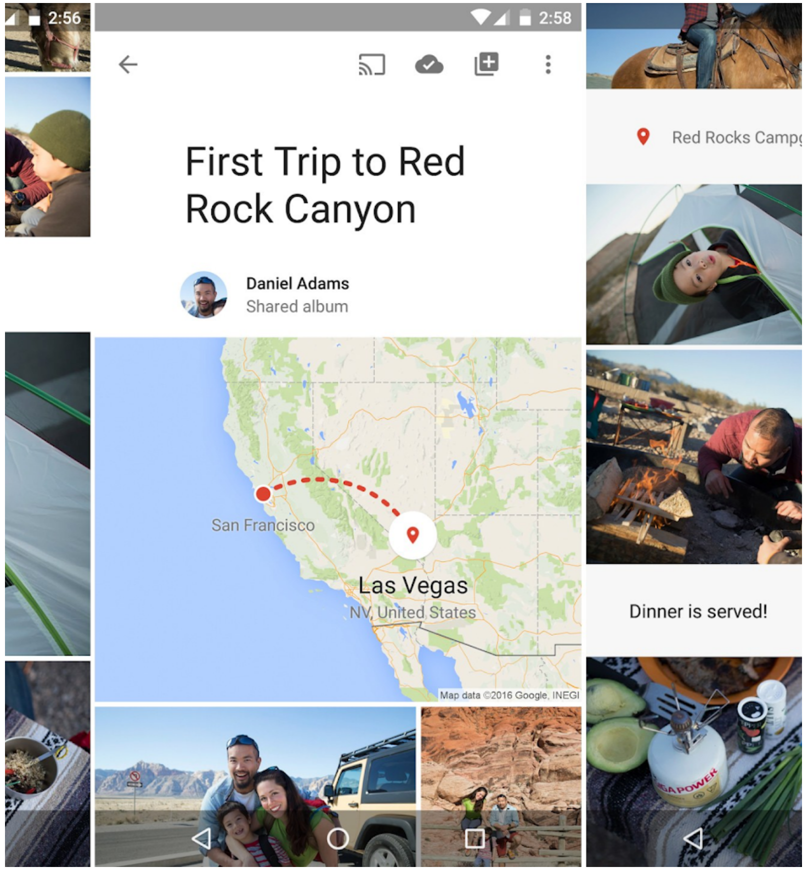
Does AI need ethics?

Teresa Scantamburlo

European Centre for Living Technology (ECLT)

Ca’ Foscari University, Venice

17 May 2018, DAIS, Ca’ Foscari University, Venice

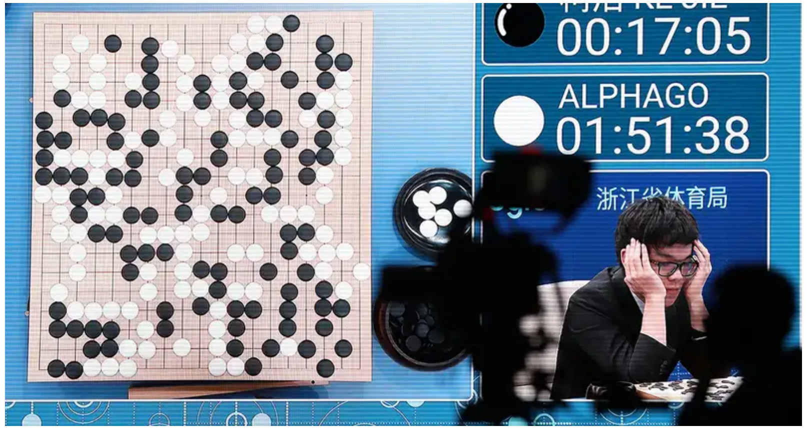


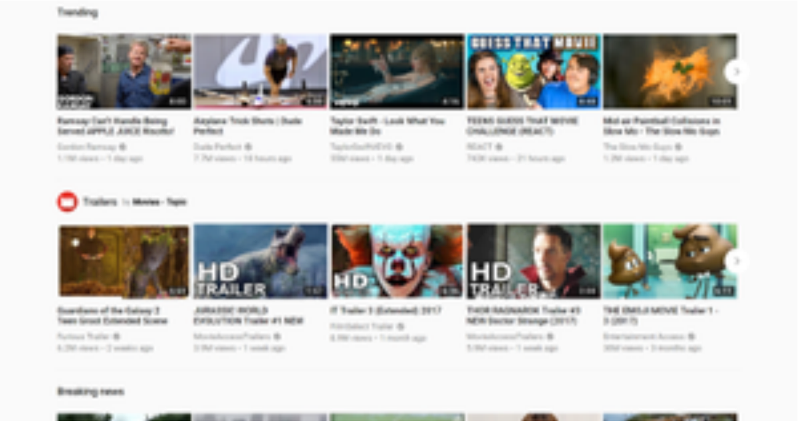
First Trip to Red Rock Canyon

Daniel Adams
Shared album

Red Rocks Campg

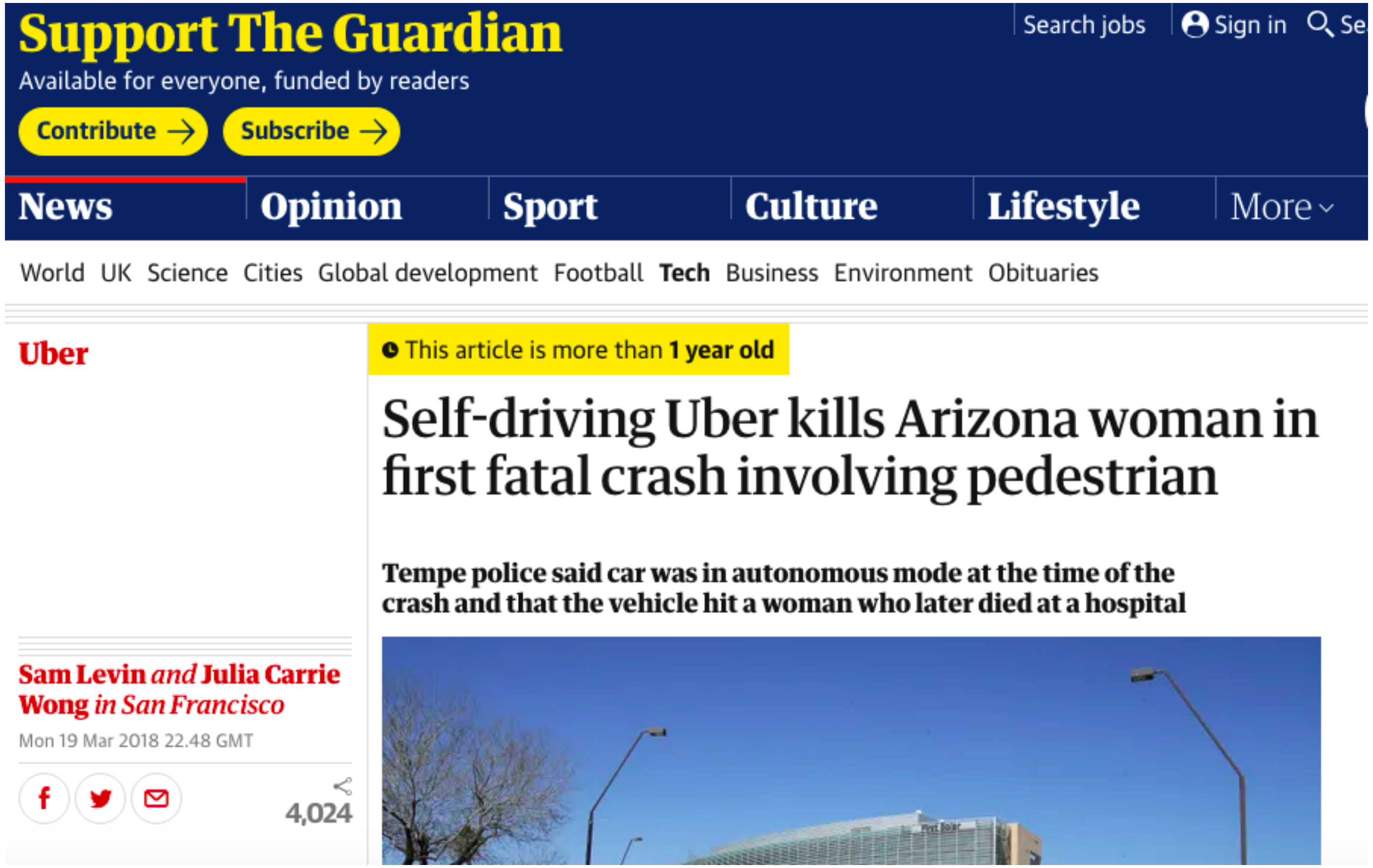
Dinner is served!





The success of AI

Uber self-driving accident



Uber

This article is more than **1 year old**

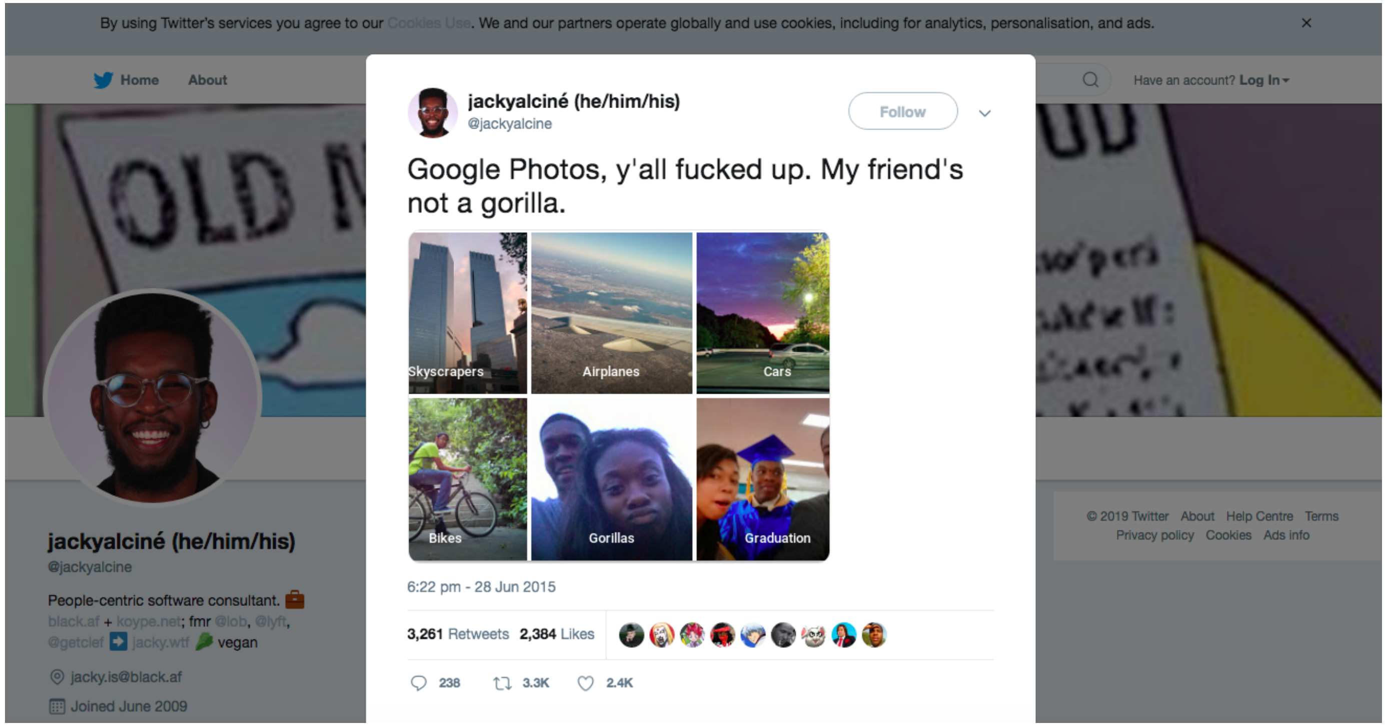
Self-driving Uber kills Arizona woman in first fatal crash involving pedestrian

Tempe police said car was in autonomous mode at the time of the crash and that the vehicle hit a woman who later died at a hospital

Sam Levin *and* **Julia Carrie Wong** *in San Francisco*

Mon 19 Mar 2018 22.48 GMT

Google photo image recognition



By using Twitter's services you agree to our Cookies Use. We and our partners operate globally and use cookies, including for analytics, personalisation, and ads.

jackyalciné (he/him/his)
@jackyalcine

Google Photos, y'all fucked up. My friend's not a gorilla.

6:22 pm - 28 Jun 2015

3,261 Retweets 2,384 Likes

238 3.3K 2.4K

jackyalciné (he/him/his)
@jackyalcine

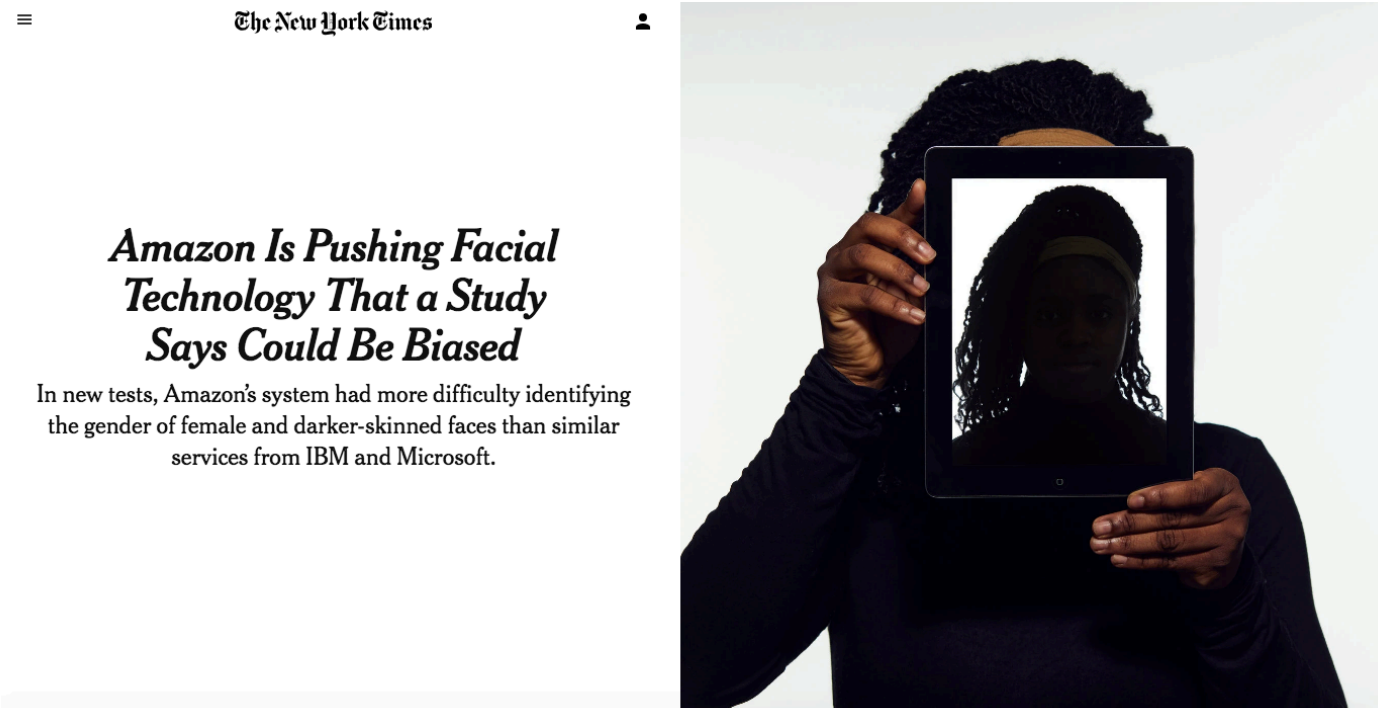
People-centric software consultant. 💼
black.af + koype.net; fmr @lob, @lyft, @getclef ➡️ jacky.wtf 🥬 vegan

jacky.is@black.af

Joined June 2009

© 2019 Twitter About Help Centre Terms
Privacy policy Cookies Ads info

Amazon facial recognition

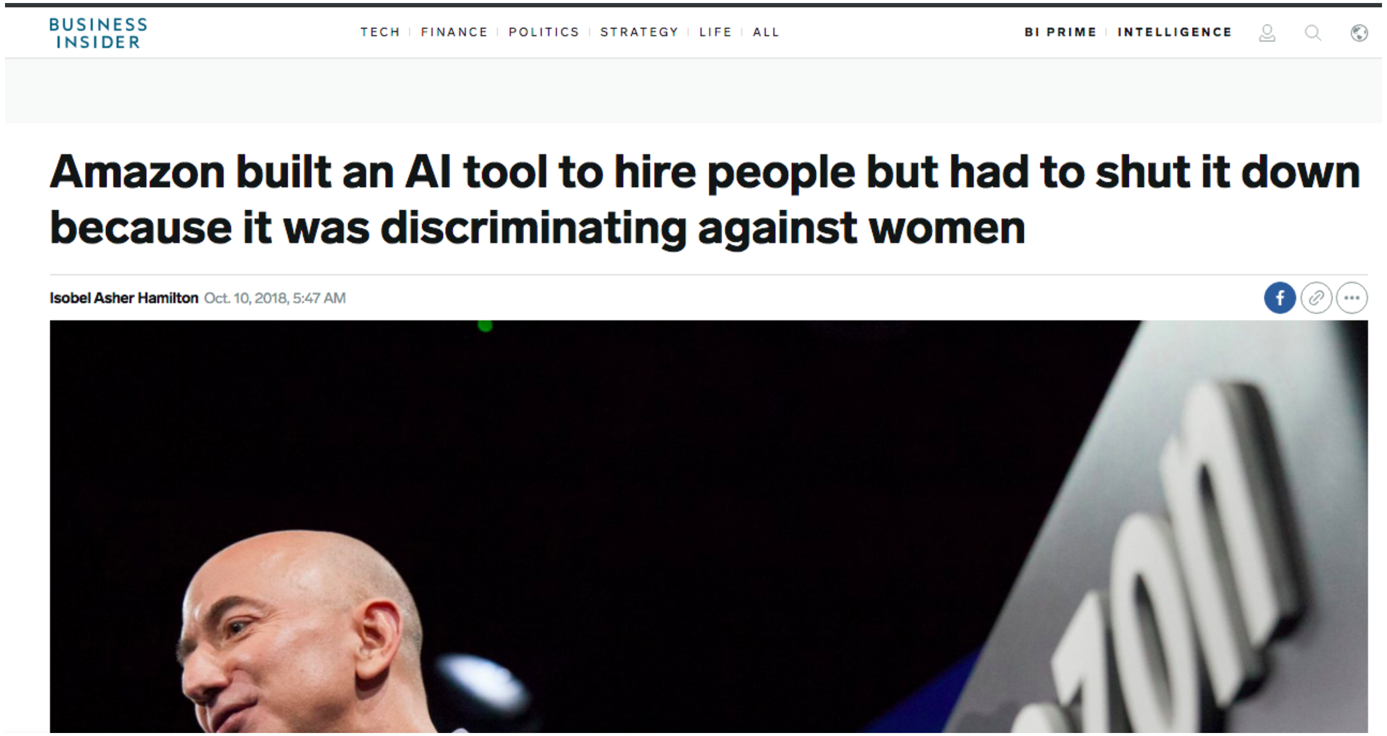


The New York Times

Amazon Is Pushing Facial Technology That a Study Says Could Be Biased

In new tests, Amazon's system had more difficulty identifying the gender of female and darker-skinned faces than similar services from IBM and Microsoft.

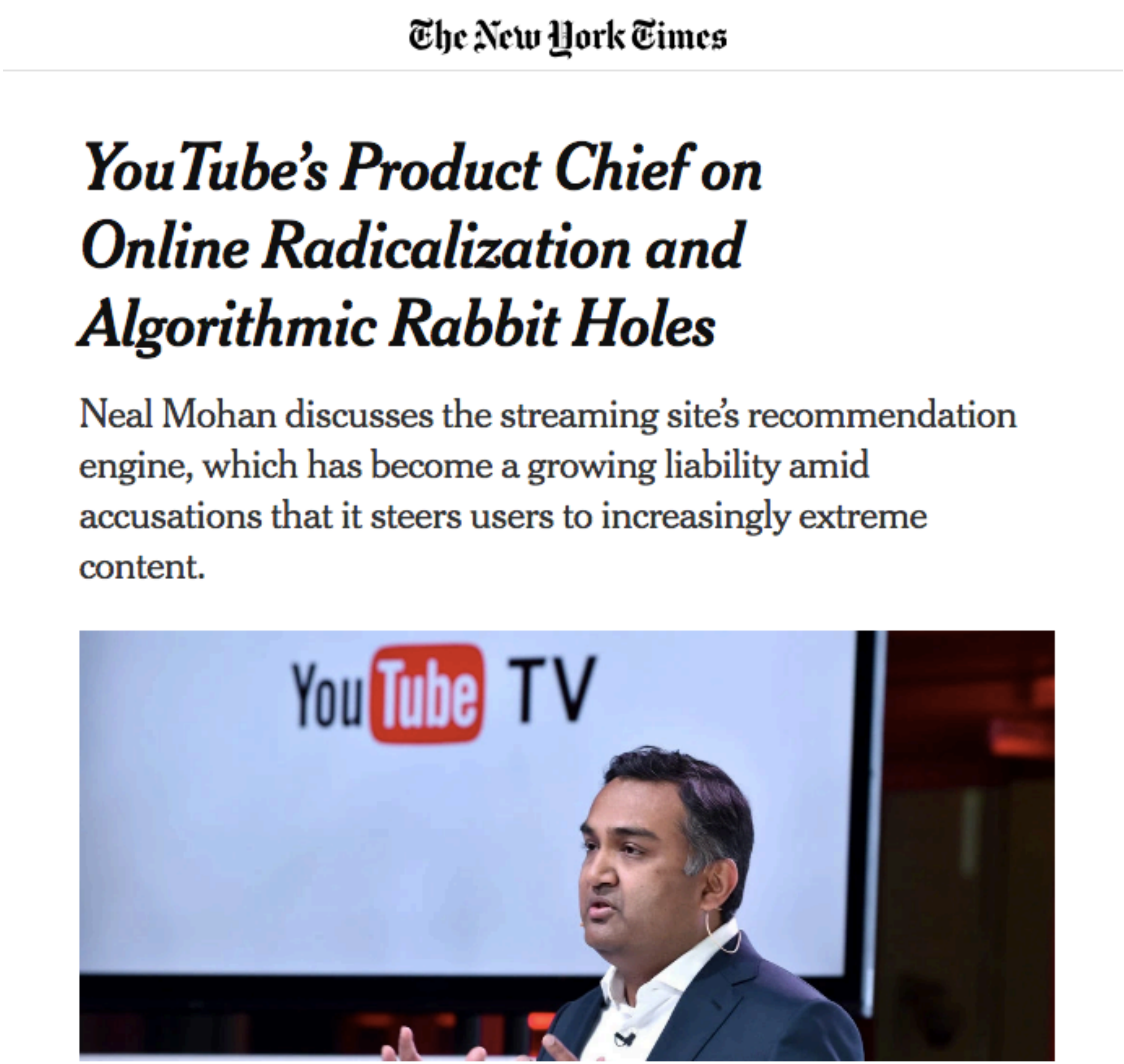
Amazon recruiting tool



Amazon built an AI tool to hire people but had to shut it down because it was discriminating against women

Isobel Asher Hamilton Oct. 10, 2018, 5:47 AM

YouTube extreme contents



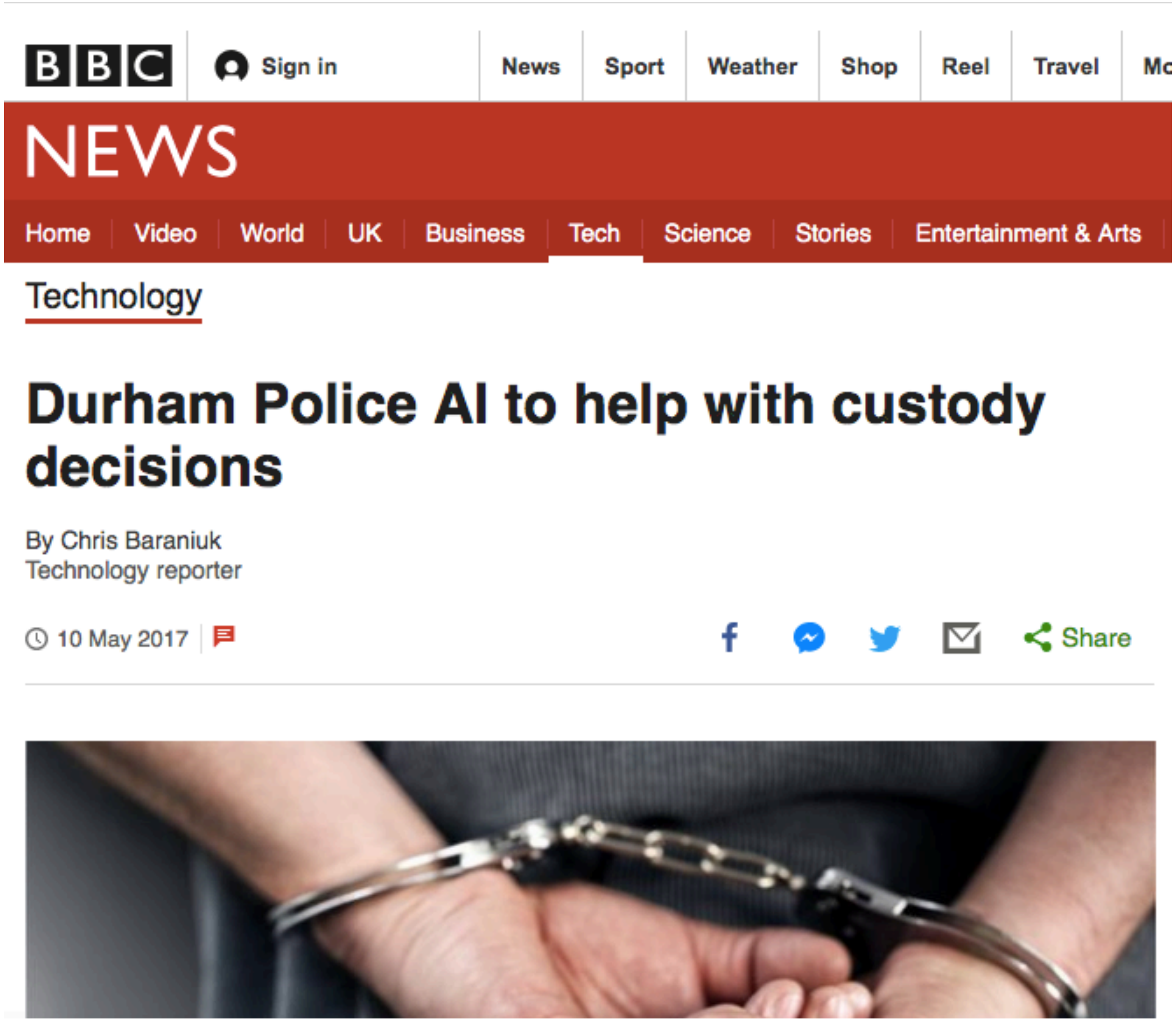
The New York Times

YouTube's Product Chief on Online Radicalization and Algorithmic Rabbit Holes

Neal Mohan discusses the streaming site's recommendation engine, which has become a growing liability amid accusations that it steers users to increasingly extreme content.

HART

Harm Assessment Risk Tool



Technology

Durham Police AI to help with custody decisions

By Chris Baraniuk
Technology reporter

10 May 2017

What is HART?

- HART = Harm Risk Assessment Tool
- It is a Risk Assessment Tool (RAT) that is used to predict the likelihood of reoffending after a follow-up period (i.e. 2 years after arrest)
- RATs are usually based on statistics or machine learning
- They can be introduced at several steps of the justice process, e.g. pre-trial hearing, early release from prison (parole), sentencing, etc.
- There are several RATs in use both in US and Europe, e.g.:
  - USA: COMPAS, Public Safety Assessment Tool, Ohio Risk Assessment System… (for a list see: https://epic.org/algorithmic-transparency/crim-justice/)
  - Europe: HART (England), OGRS (England and Wales), StatRec (Netherland), Static99 (just for sexual offenders, Netherland)

Our sources

- The analysis of HART is based on the following sources:
  - Urwin S (2016) *Algorithmic Forecasting of Offender Dangerousness for Police Custody Officers: An Assessment of Accuracy for the Durham Constabulary*. Master Degree Thesis, Cambridge University, UK
  - Oswald M, Grace J, Urwin S and Barnes GC (2018) Algorithmic risk assessment policing models: lessons from the Durham HART model and 'Experimental' proportionality, *Information & Communications Technology Law*, 27(2): 223-250
  - Barnes G, and Hyatt J (2012) , Classifying Adult Probationers by Forecasting, Future Offending, Tech report
- Extensive media coverage
  - BBC: https://www.bbc.com/news/technology-39857645

A brief sketch

- Launched in May 2017
- Developed by Durham Police in collaboration with Cambridge University
- **Objective**: to support custody decision = “decision taken by the custody officer following arrest at the end of the first custody period” (Urwin, 2016)
- Model’s **output**: “high-risk” – “moderate-risk” – “low-risk”
- Context: “checkpoint programme” that aims at providing “moderate-risk offender” with an alternative to prosecution (https://www.durham.police.uk/Information-and-advice/Pages/Checkpoint.aspx)

HART’s training set

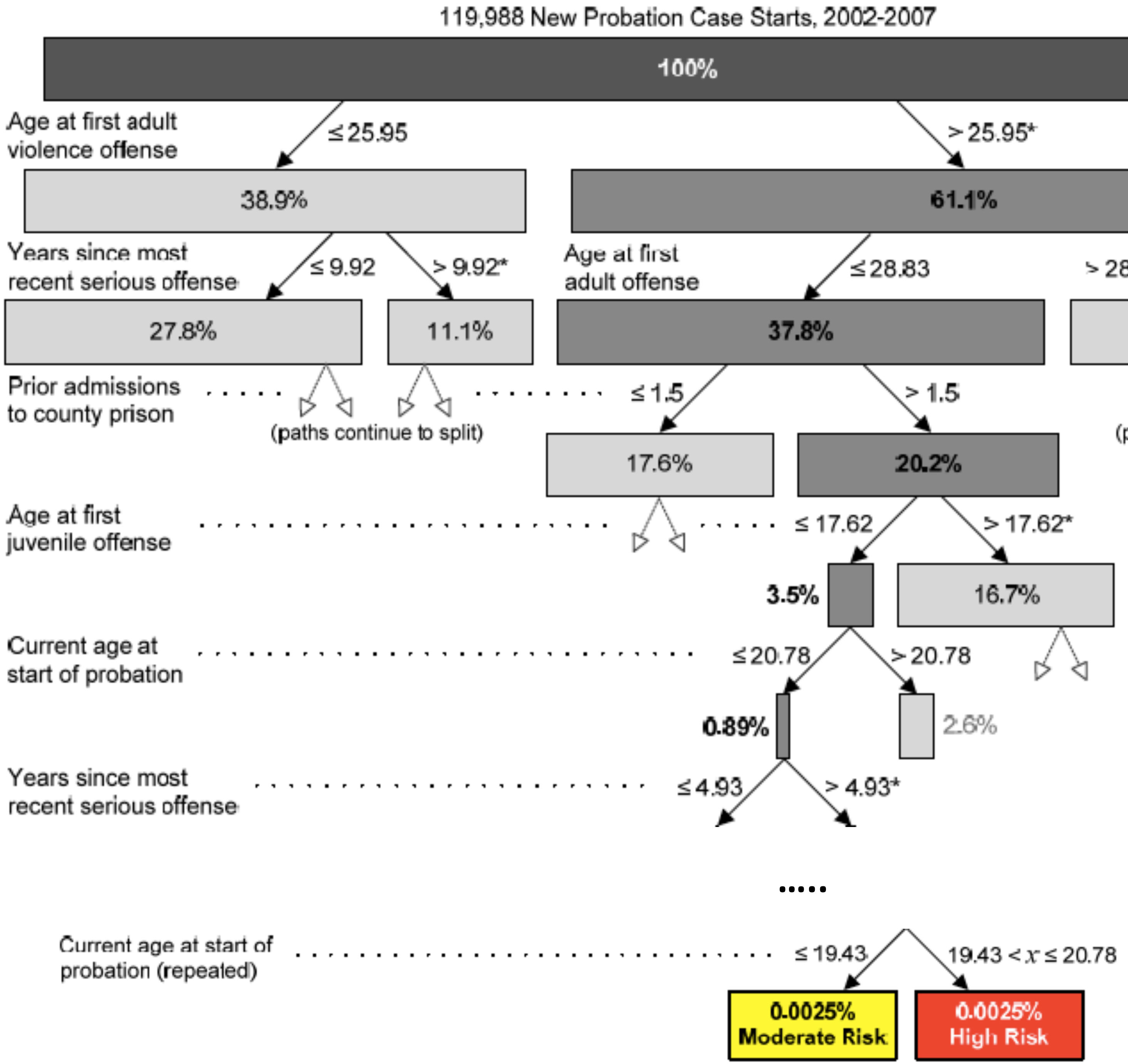
- 104,000 custody events within a period between Jan 2008 and Dec 2012

- 34 features such as:
  - Age at custody event
  - Gender
  - Count of any past offences
  - Instant violence offence (Y/N)
  - Custody Outward Postcode (3-4 first characters)
  - (Experian) Mosaic Code (socio-geo demographic)
  - Age at first offence
  - …

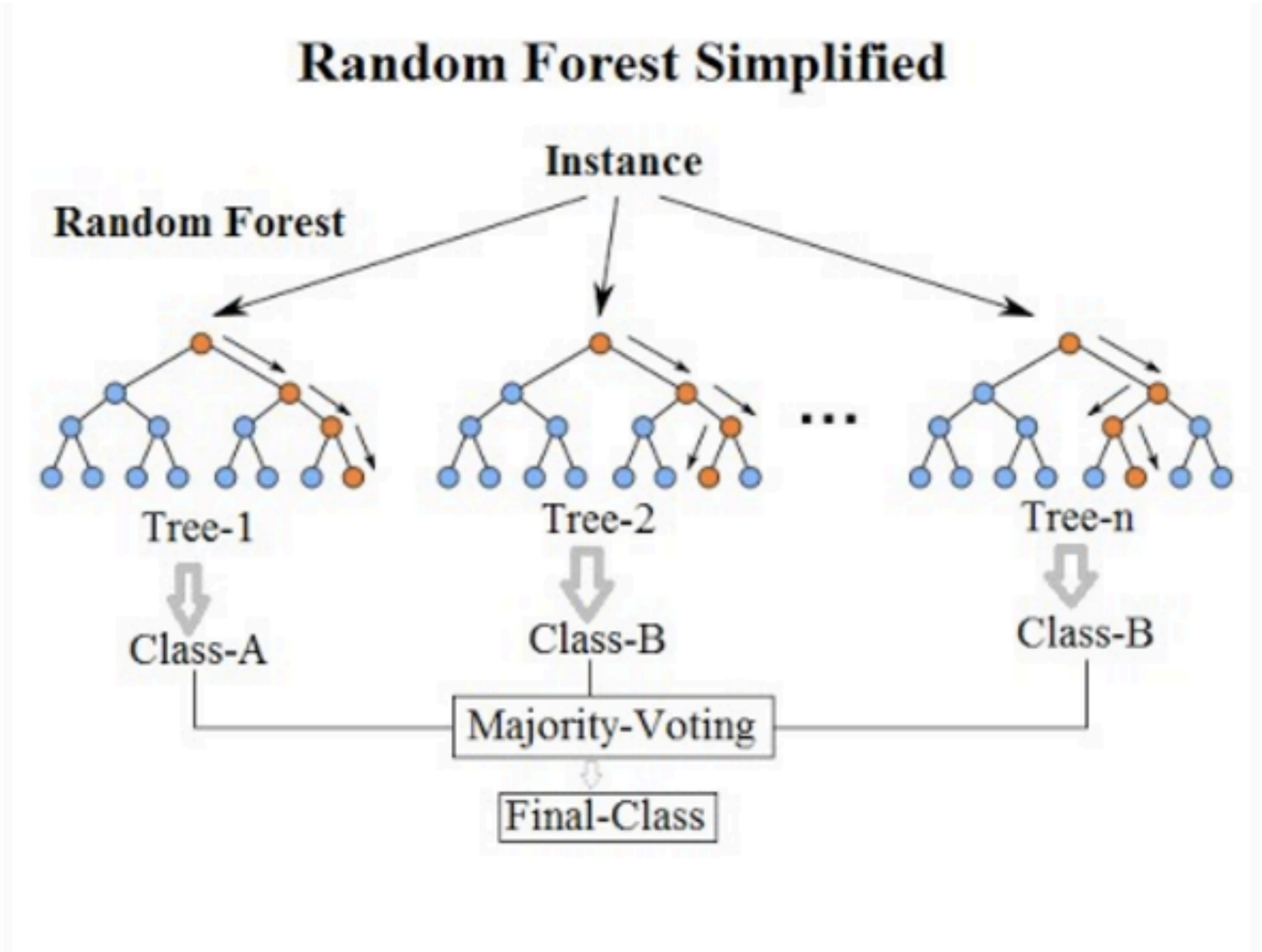
- Categorical labels:
  - High-risk = a new serious offence within the next 2 years
  - Moderate-risk = a non-serious offence within the next 2 years
  - Low-risk = no offence within the next 2 years

HART’s model

- HART is based on Random Forest, a ML method that results from the combination of a multitude of decision trees
- A decision tree is a popular classification technique that tests an attribute at each node and assign instances to the descending branches based on the value taken by instances for that attribute
- Each decision tree is trained on a random subsamples of the training set and using a random subset of features
- HART uses 509 decision trees, each producing a prediction. The output corresponds to the output that receives the most votes


119,988 New Probation Case Starts, 2002-2007
100%
Age at first adult violence offense
≤ 25.95
> 25.95*
38.9%
61.1%
Years since most recent serious offense
≤ 9.92
> 9.92*
Age at first adult offense
≤ 28.83
> 28.83*
27.8%
11.1%
37.8%
23.3%
Prior admissions to county prison
(paths continue to split)
≤ 1.5
> 1.5
(paths continue to split)
17.6%
20.2%
Age at first juvenile offense
≤ 17.62
> 17.62*
3.5%
16.7%
Current age at start of probation
≤ 20.78
> 20.78
0.89%
2.6%
Years since most recent serious offense
≤ 4.93
> 4.93*
.....
Current age at start of probation (repeated)
≤ 19.43
19.43 < x ≤ 20.78
0.0025% Moderate Risk
0.0025% High Risk

Barnes G, and Hyatt J (2012) , Classifying Adult Probationers by Forecasting, Future Offending, Tech report



https://commons.wikimedia.org/wiki/File:Random_forest_diagram_complete.png

Technical details

- Out-of-bag error = when a random sample is drawn to grow a decision tree a small amount is held out and used as a test set to estimate the generalization error during training

- Weighting different types of errors:
  - **Dangerous errors**: misclassifying a serious offender as a low-risk
  - **Cautious errors**: misclassifying a non-serious offender as a high-risk

- Policy decision: HART weights more dangerous error (i.e. it applies a lower cost-ratio)

Performance measures

Comparison with the accuracy of a random baseline:

[P(Y = “high”) * P(Ŷ= “high”)] + [P(Y = “moderate”) * P(Ŷ= “moderate”)] + [P(Y = “low”) * P(Ŷ= “low”)] =

[0.1186 * 0.1186] + [0.4835 * 0.4835] + [0.3979 * 0.3979] = 0.406 = **41%**

	OOB construction data	2013 validation data	
Overall accuracy: what is the estimated probability of a correct classification?	68.50%	62.80%	
Sensitivity / recall: what is the true positive rate for each class label?	72.60%	52.75%	HIGH
	70.20%	67.28%	MODERATE
	65.30%	60.35%	LOW
Precision: what is the rate of relevant instance for each class label?	48.50%	33.83%	HIGH
	70.20%	63.84%	MODERATE
	75.60%	78.60%	LOW
Very dangerous errors: of those predicted low risk, the percent that was actually high risk (subset of the false omission rate)	2.40%	2.38%	
Very cautious errors: of those predicted high risk, the percent that was actually low risk (subset of the false discovery rate)	10.80%	12.06%	

some performance measures of HART extracted from tables 6 and 9 in Urwin (2016: 52,56)

Is Accuracy enough?

From accuracy to trust

- Being accurate is not enough
  - What performance measures are used?
  - What sample is used for validation?
- How are decisions made?
- What model has been used? What features?
- Can we explain the logics behind the algorithm to the interested subject? (GDPR)
- Is the algorithm fair? Or does it discriminate?
- …

Transparency

- The ability to access information about the procedures and the data that lead to a certain decision (e.g. school admission or access to credit)
- Desirable property of legal and administrative process
- It helps an agent (e.g. a bureaucracy) be accountable for the performed task: tracking the whole process, making information available for (external) audit, allowing citizen to question or contest a decision
- Transparency is not sufficient:
  - Trade secret
  - Disclosure of sensitive information
  - Gaming the system
  - Black-box model (non understandable by humans)

“Right to explanation”

- EU’s General Data Protection Regulation came into effect in May 2018
- GDPR specifies various obligations for the data controller concerning data collection and processing
- In article 13(2) it states that the data controller should provide the data subject with the following information:

> the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject

Example

- Consider our case study:
    - How would you “explain” the output of HART?
    - How would you assess the role played by critical features such as the postcode or Experian code?
- Objection: also humans can provide poor explanations
- Explanation is a controversial notion: we don’t require a judge to report his/her mental states that cause a certain decision. We ask for justification
- But, can you interrogate an algorithm?

Algorithmic fairness

- Ethical/legal principle = “all human beings must be treated equally before the law” (human rights declarations)
- No universal definition (philosophy / philosophy of law)
- Typically, regulations identify a list of attribute characterizing protected groups and with respect we don’t want to discriminate, e.g.:
  - Race
  - Color
  - Sex
  - Religion
  - language
  - Sexual orientation
  - Political opinion
  - …

For a complete list see, e.g., the *International Covenant for Civil and Political Rights* (article 26) and the *European Convention on Human Rights* (article 14)

Practical goal: to avoid a different / adverse treatment based on any of such attributes

The source of algorithmic discrimination

- How does an algorithm can discriminate?

  - Feature selection: when using sensitive attribute that directly or indirectly correlates with the membership to a protected group
  - *Example*: using a sensitive attribute or a feature that may correlate with membership to a protected group. See the use of postcode or Experian code in HART

  - Sampling: when the training set incorporates bias
  - *Example*: language is full of bias (see the work by Caliskan et al. 2017) and, in case of law enforcement, also judges' or police officers' decisions, which are used as training examples, can be biased and based on prejudices.

  - Unbalances in the training sample
  - *Example*: we may over/under-represent some social groups so that the model learns badly for certain categories. See the poor performance of certain facial/image recognition systems when tested over specific categories

Operationalizing fairness

- Techniques for discrimination discovery and discrimination prevention (e.g. during pre-processing or post-processing). For a survey see Zliobaite, 2017
- Fairness criteria for quantifying unfairness in classification
- Formal set-up:
    - P (Y,X,A) = joint probability distribution
    - Y = outcome of interest, e.g. Y={1, 0}
    - X = set of “legitimate features”
    - A = protected attributes, e.g. A={‘f’, ‘m’}
    - $\hat{Y}$ = f(X, A)

Note that the assessment of fairness criteria needs information about the size of the protected groups within the dataset.

Fairness criteria

- Some examples include:
- Statistical parity:
  - $P(\hat{Y} = 1 \mid A = \text{‘f’}) = P(\hat{Y} = 1 \mid A = \text{‘m’})$
  - Example: fraction of people classified as “high-risk” is equal across the groups
- Calibration:
  - $P(Y = 1 \mid \hat{Y} = 1, A = \text{‘f’}) = P(Y = 1 \mid \hat{Y} = 1, A = \text{‘m’})$
  - Example: fraction of people correctly classified as “high-risk” is equal across the groups
- Error rate balance:
  - $P(\hat{Y} = 1 \mid Y= 0, A = \text{‘f’}) = P(\hat{Y} = 1 \mid Y = 0, A = \text{‘m’})$ and $P(\hat{Y} = 0 \mid Y= 1, A = \text{‘f’}) = P(\hat{Y} = 0 \mid Y = 1, A = \text{‘m’})$
  - Example: false positive and false negative rates are the same across the groups

Racial discrimination in COMPAS

ProPublica's analysis suggested that COMPAS had problem with **error rate balance**:

- black defendants were nearly twice as likely to be misclassified as higher risk compared to white defendants (45% vs. 23%)
- white defendants who re-offended within next 2 years were mistakenly labelled "low-risk" twice as often as black re-offenders (48% vs. 28%)

Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

Impossibility result

- The bad news is that fairness criteria cannot be simultaneously satisfied
- Example:
  - ProPublica’s analysis found that COMPAS doesn’t satisfy the error rate balance
  - Another independent analysis showed that COMPAS satisfied calibration (Flores et al, 2016)
  - Northpointe’s showed that COMPAS satisfied predictive parity (similar to calibration)
- Some references on the impossibility result
  - Chouldechova A (2017) Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, arXiv
  - Kleinberg J, Mullainathan S & Raghavan M (2017) Inherent trade-offs in the fair determination of risk scores, *(ITCS 2017)*
  - Berk R, Heidari H, Jabbari S, Kearns M & Roth A (2017) Fairness in Criminal Justice Risk Assessments: The State of the Art, *Sociological Methods & Research*

Beyond technical solutions

- Ethics is about good and bad choices/actions.
- Many of these choices can be translated into technical constraints
- But technical remedies requires human discernment, the capacity of reflecting on situations (persons involved, values, consequences, risks, benefits,) and choosing responsibly
- There is no recipe that guarantees the success of human discernment.
- Practical wisdom combines several rational, emotional and social skills
- Education and interdisciplinary dialogue are a powerful means

Applied ethics

Some ideas to outline a path:

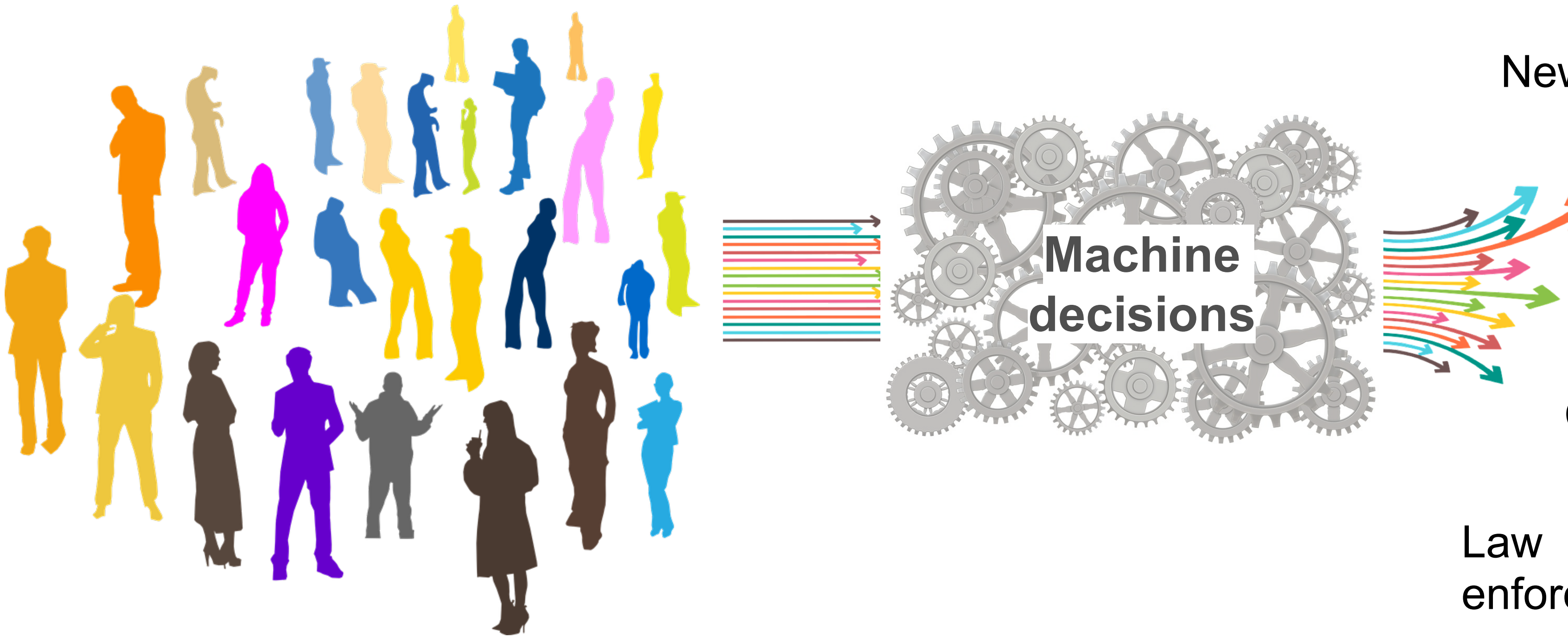
- Doing your job properly and being aware of your obligations
- Identifying the social and ethical issues in your field
- Analyzing problems in terms of facts, values, stakeholders, impact..
- Confronting with ethical frameworks (e.g. Kantian theory, utilitarianism, virtue ethics, etc.), professional ethics and relevant regulations
- Making judgments, creating new course of actions and providing justifications
- Discussing with collogues and people with different background
- …

Some concluding remarks

- The (alleged) promises of algorithms
  - Reduction of costs/resources
  - Neutrality of the tools (i.e. unbiased)
  - Objectivity and precision of judgement
  - …
- Many of these assumptions turned out to be untrue (algorithms are not neutral!)

- In any case, these algorithms have a direct impact on human lives and they should be carefully assessed

Algorithms as new gatekeepers



Machine decisions

School

News

University

Credit

Social care

Books

Health care

Clothes

Music

Job

Law enforcement

Videos

Food

Movies

Classifying humans is not the same as classifying objects

Scantamburlo T., Charlesworth A. and Cristianini N. (2019). "Machine decisions and human consequences". In K. Yeung & M. Lodge (eds) Algorithmic Regulation (forthcoming). Oxford: Oxford University Press (available on arXiv)

Feedbacks, comments or requests are welcome
teresa.scantamburlo@unive.it

Thanks!