CATÓLICA
FACULTY
OF LAW

ESCOLA DO PORTO

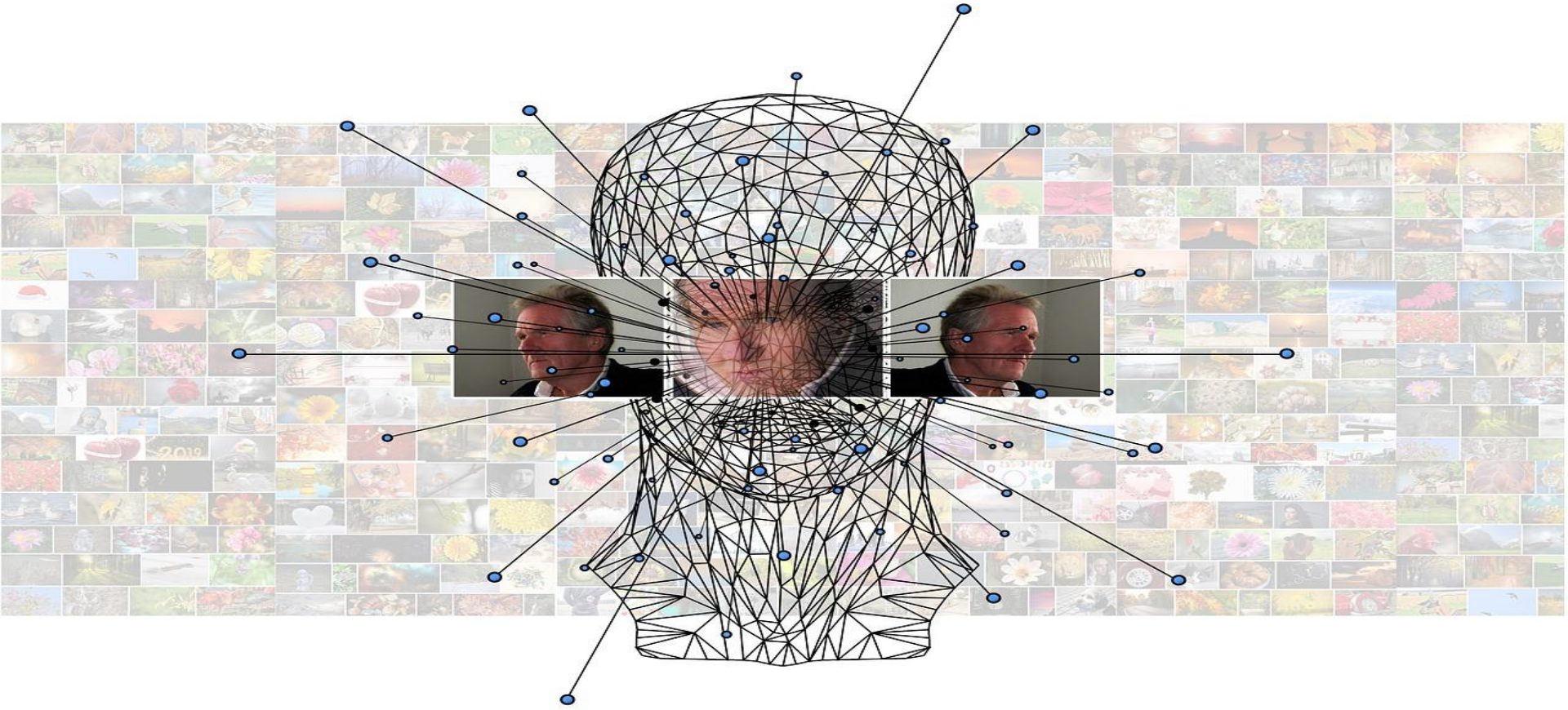
Funded by
the European Union

The end of the deception? – Counteracting algorithmic discrimination in the digital age

Catarina Santos Botelho
cbotelho@ucp.pt

What are algorithms?

- Algorithms are encoded procedures through which input data, following a logical and mathematical sequence, transform into an output.



- At this stage, algorithms are complex and defy human comprehension.
- Algorithms can be categorized in knowledge-based and machine-learning (ML) algorithms.
- **Knowledge-based** systems are the traditional flowchart algorithms that require manual coding.
- **ML** are algorithms that feed large amounts of data with output variables for the algorithm to “learn without being explicitly programmed”.

Why is this topic important?

Automated decision-making processes are an increasing reality

- healthcare provisions,
- vaccine allocation,
- prison sentences,
- police scrutiny,
- IRS audit selection,
- vote counting,
- loan determinations,
- job applications,
- work allocation,
- social benefits requests,
- granting of immigration visas,
- kindergarten or university admission selections,
- prices,
- consumer surveillance,
- micro-targeted marketing.

What are the big questions?

- What's next for democracy and human rights protection in a newfound "socio-technical Internet architecture"?
- Can algorithms intensify democratic backlashing?
- Is human agency in decision-making processes still vital for human rights protection or an unnecessary precaution?

Direct discrimination / indirect discrimination

- **Direct**: when a person is treated less favourably than another because of a protected characteristic in matters of a protected sector.

- **Indirect**: when a seemingly neutral provision, criterion or practice puts a person of a protected group at disadvantage without an acceptable justification.

Disparate treatment / Disparate impact

- Title VII of the Civil Rights Act of 1964 (USA)

- *Disparate treatment:* when the employer treats the employee differently because of a protected characteristic.
- *Disparate impact:* a practice which on its face appears neutral and non-discriminatory, but in fact disproportionately disadvantages a protected group.

Protected groups criteria

- Are protected groups fixed or can other “socially salient axes of identity” be added?
- Can the socio-economic status or the education level be considered for anti-discrimination proposes? And what about obesity, chronic sickness, or browser type?
- Should anti-discrimination law design, in particular the lists of ground and/or exceptions, be open, closed, or hybrid?

Biases regarding ML algorithms:

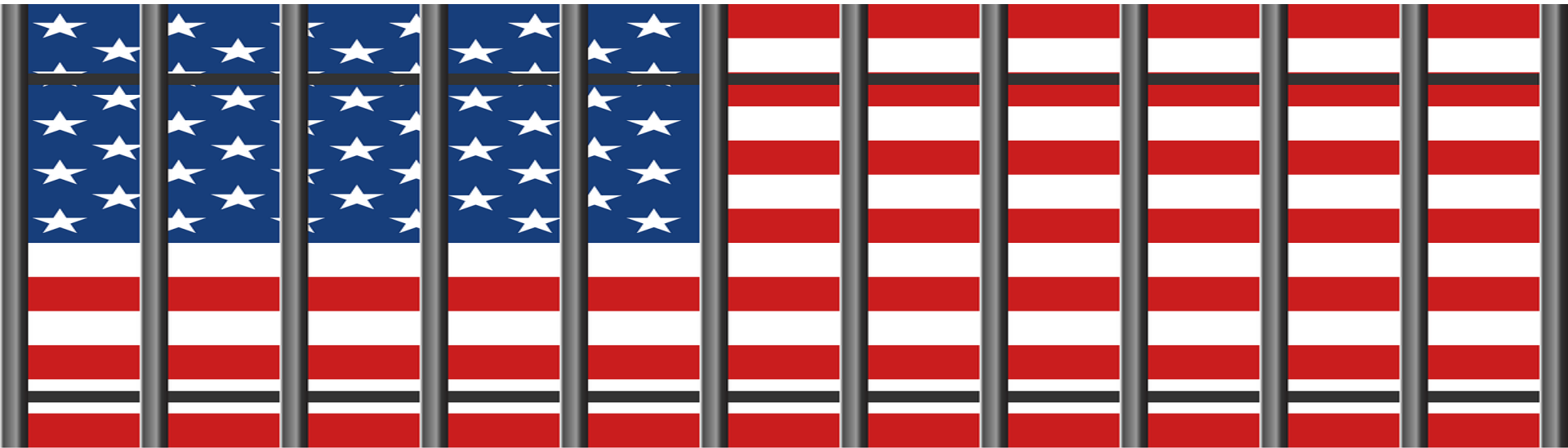
- *consequential bias*, when the creators of the algorithm are intentionally or unintentionally biased, and that bias will, thus, replicate in the data-processing mechanism;
- *biased sample data*, in which the provided sample or training data are biased;
- *outcome bias* or *vicious circle bias*, which happens even when the algorithm is trained with representative data, due to the data's embedded social inequalities.

Examples of algorithmic disparate impact and statistical discrimination

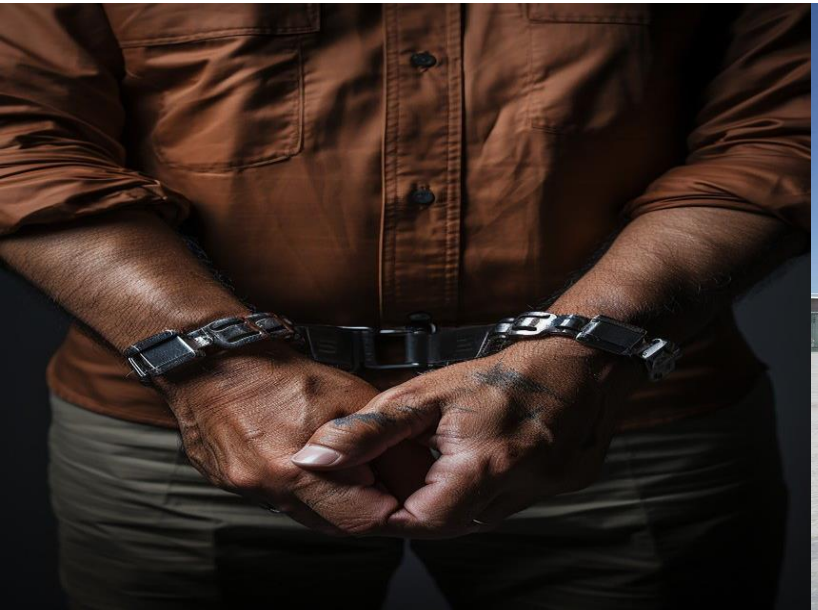
- ***Amazon case***:
- Amazon developed a ML system to identify *future software developers' employees*. The algorithm ranked job candidates on expected productivity. As the algorithm was trained using existing hiring data – which reflected the existing male dominance in the technology industry – candidates with female indicators on their resumes (e.g., women's tennis club) were discriminated regardless of their qualifications for the job.

- ***Correctional Offender Management Profiling for Alternative Sanctions*** **(COMPAS) case:**

- Using data from a USA county, a company developed a risk assessment tool known as COMPAS to predict recidivism for detainees and defendants.
- For the question “will this person recidivate if given parole?” the algorithm compared the person’s characteristics with the characteristics of those that recidivated during parole.



- The problem that ProPublica reported in 2016 was that COMPAS employed its prediction of re-arrest as a proxy for re-offence.
- COMPAS was accused of yielding racially biased results, since false positives (foreseeing a hight risk of recidivism where the offender does not actually reoffend) for African-American defendants doubled false positives for Caucasian origin defendants.



The illusion of algorithmic objectivity

- “Algorithmic aversion” / “AI-as-the-Monster”

versus

- Herculean super-powers / champions of objectivity

Endogenous algorithmic discrimination

- Discrimination lies in the algorithmic design itself.
- More transparency: prevent migration of discriminatory human decision-making to algorithm decision-making.
- Is transparency the new consent trap?

Exogenous algorithmic discrimination

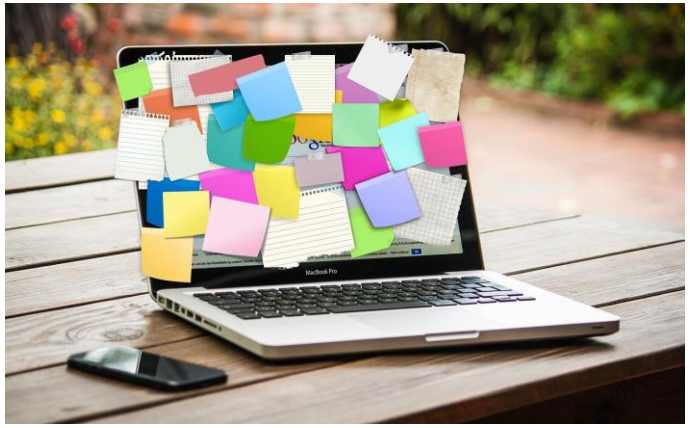
- The problem is not the 'black box' itself, but the actual *world* in which algorithms operate.
- Discrimination is a result of implicit biases, statistical distortion, or historical group misrepresentations that made their way through supposedly aseptic ML.
- The idea is to "move beyond merely *non*discriminatory to actively *anti*discriminatory algorithms" (Bornstein).

Substantive equality

- ***Algorithmic affirmative action*** can be achieved through several measures, such as incorporating equity metrics into the model selection processes, or through group fairness using “demographic parity” or “statistical parity” approaches.

Regulating and debiasing algorithms

- Since algorithms are incomprehensible to humans, an additional endeavour is needed to discover discriminatory outputs.
- Under these circumstances, policymakers, courts, agencies, public authorities, and employers should not adopt a romanticized stance of uncritically and excessively trusting algorithms.
- An active vigilance is, thus, highly recommended.



1. Algorithmic control

- The European AI Act pioneeringly aims at a human-centric and ethical AI, in which AI system's outputs are traceable, explainable, and transparent.

- AI systems are categorised based on their level of potential risk (Article 5):

- Prohibited a priori
- High-risk
- Low-risk

- Proportionality test.

- To minimize the potential impact of high-risk AI systems, deployers need to conduct **Fundamental Rights Impact Assessments (FRIA)** prior to deploying a high-risk AI system (Article 27).

Article 6 (European AI Act)

*3. By derogation from paragraph 2, an AI system shall not be considered to be high-risk if it does not pose a **significant risk** of harm to the health, safety or fundamental rights of natural persons, including by not materially influencing the outcome of decision making. This shall be the case where one or more of the following conditions are fulfilled:*

(a) the AI system is intended to perform a narrow procedural task;

(b) the AI system is intended to improve the result of a previously completed human activity;

(c) the AI system is intended to detect decision-making patterns or deviations from prior decision-making patterns and is not meant to replace or influence the previously completed human assessment, without proper human review; or

(d) the AI system is intended to perform a preparatory task to an assessment relevant for the purposes of the use cases listed in Annex III.

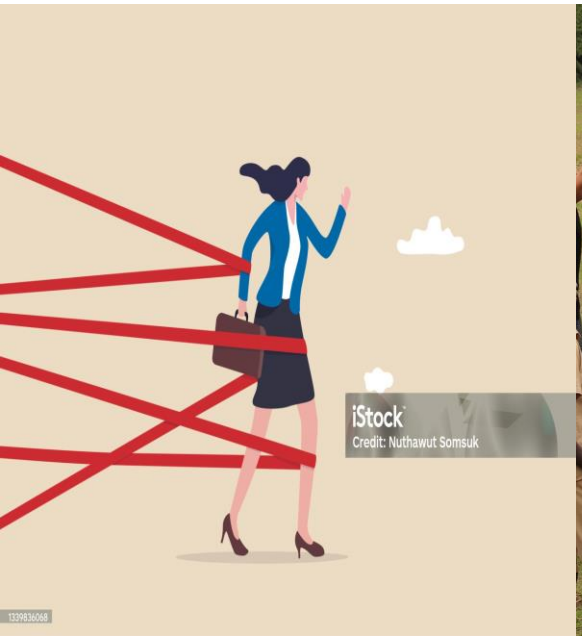
2. Blocking information: “fairness through blindness?”

- Does blocking personal information about protected categories in training data prevent discriminatory outcomes?
- To deliver race-blind, disability-blind, gender-blind, income-bind, or zip-code-blind results, it may not be sufficient (nor even wise) to simply remove sensitive attributes from consideration.

- ML systems excel at “detecting patterns, and at identifying and leveraging on nonobvious proxies for omitted sensitive variables”.

- How can we disclose “layers upon layers of mirrors and proxies”, and how do we know which *proxies* (functional substitutes) contain valuable information?

- An additional complicating factor is that parity regarding race and gender may still neglect *intersectional differences* within those groups along other dimensions, such as disabilities, age, or nationality.



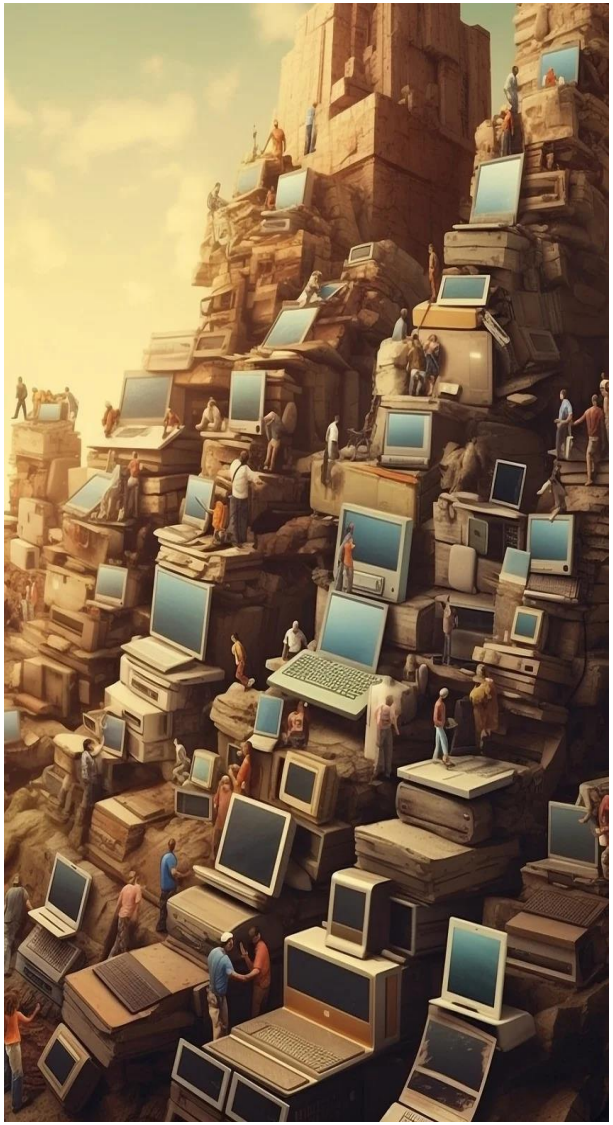
3. Regulating and shaping data through algorithmic design

- The General Data Protection Regulation (GDPR) requires data processing activities to be transparent (Recitals 39, 58, 78 GDPR and Articles 5 § 1 (a) and 12 GDPR), grants the right to access and the right to be informed of the existence of automated decision-making (Articles 13-15), as well as the right not to be solely subject to automated decision-making (Article 22), and introduced the data protection impact assessment, to assess the risks to rights and freedoms of data subjects and the risks of the processing activity (Article 35). However, this right does not apply when the individual has provided explicit consent for the automated decision (Article 22 § 2 (b) GDPR).

At this stage, more important than regulating algorithms is regulating the data that is fed to the algorithms.

The aim is gathering "*fair* data" that can counterbalance societal biases.

Instead of blocking the data, some scholars propose *altering* them. Pre-processing data might be achieved by adjusting the input data generally or by adjusting the target variable (e.g., gender).



4. Ex post strategies for repairing discrimination: individual rights or group rights?

- courts (via judicial review)
- data protection authorities (e.g., through data protection impact assessments)
- “FDA for algorithms”?
- “an individual right to contest AI decisions”?

Algorithms' blind spots and the importance of human-centredness

- Is there "a right to a human decision"?
- AI can mimic and even surpass relevant parts of the human intellect. Nevertheless, human inductive reasoning, human cognition, and the human presence in exercising case-by-case *discretion* (which is different than arbitrariness) or judgment are still necessary.

Why?

- For the reason that only the human person can fully apprehend the reality. AI executes logical operations that are abstractions from reality (the outside world or the real world). By doing so, AI creates its reality within ‘the reality’, although it is always a narrow reality that cannot ontologically reach its *essence*. On the contrary, humans *experience and embrace reality*.

- This is not a futile competition between humans and machines. We do not need human intervention in AI because ‘we, humans’ are better.
- We need human guidance because we cannot skip humanity, ethics, and morals from public and private decision-making processes that might deeply affect our lives.