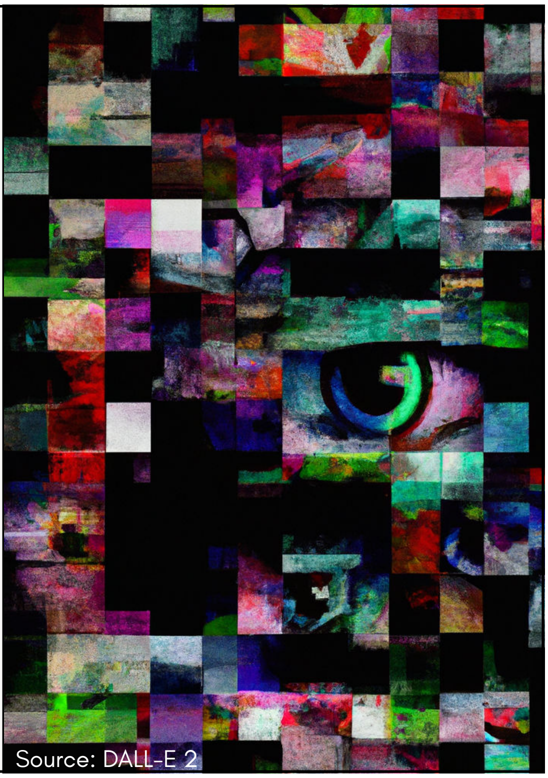
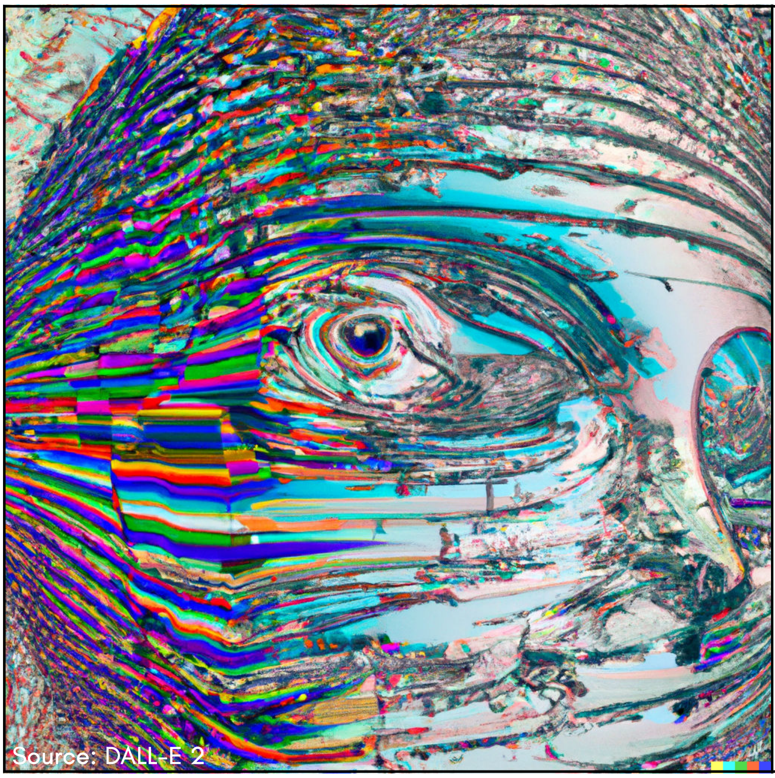
When computers say no: Addressing algorithmic bias through EU anti-discrimination law

Dr Raphaële Xenidis, Sciences Po, Law School (Paris, France)

Funded by the European Union



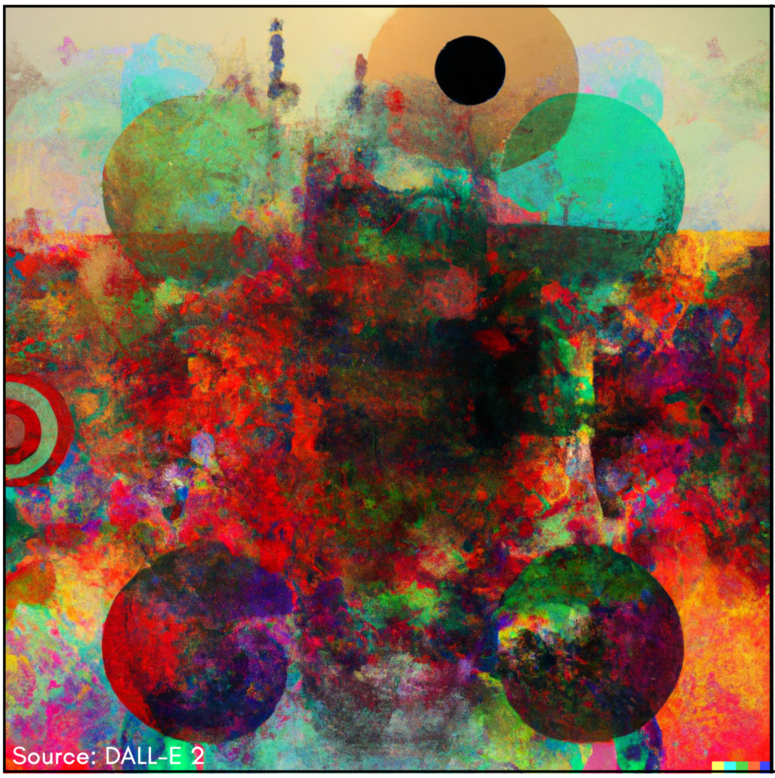
Source: DALL-E 2



Source: DALL-E 2

Outline

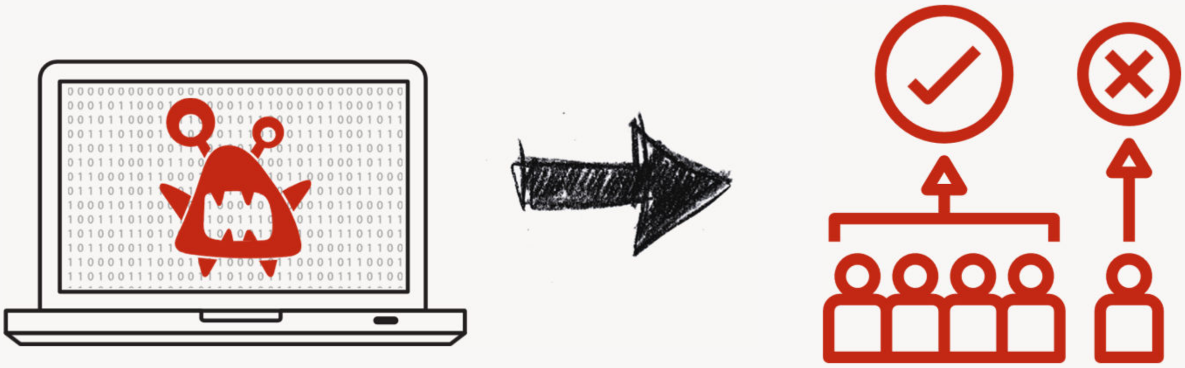
- How do machines discriminate?
- What is discrimination in EU law?
- What are the legal remedies against algorithmic discrimination?
- The interaction between EU anti-discrimination law, data protection and AI regulation: some reflections
- Shortcomings and remaining issues



Source: DALL-E 2

1. How do algorithmic machines discriminate?

The problem



Bias

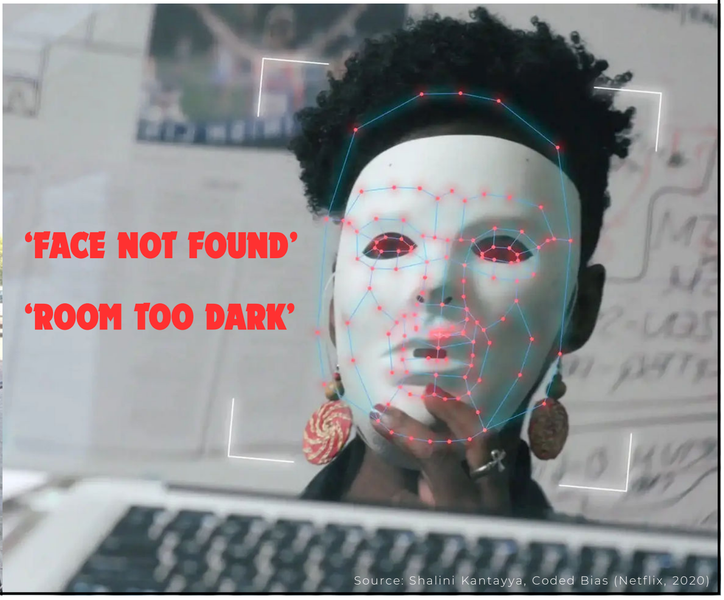
Discrimination

Photo by Jonathan Borba on Unsplash

AI'S BROKEN PROMISES

VIGNETTE I: FACIAL RECOGNITION IN EDUCATION

Source: https://racismandtechnology.center



Source: Shalini Kantayya, Coded Bias (Netflix, 2020)

AI'S BROKEN PROMISES

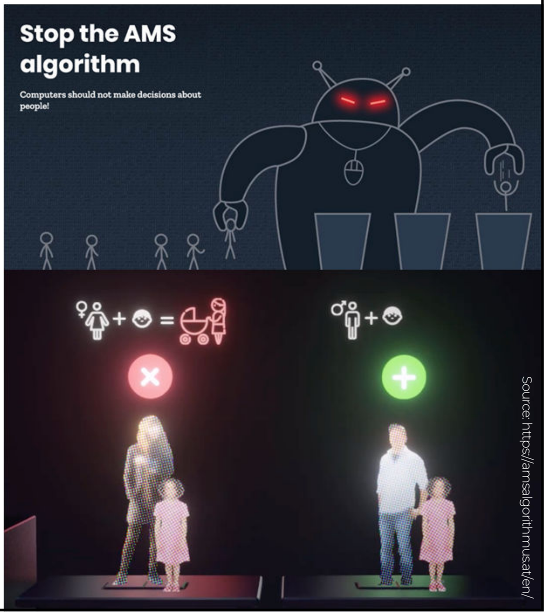
VIGNETTE II : RISK SCORING IN THE EMPLOYMENT MARKET

Variables used in the satistical model

Variable	Nominal values
Gender	Male/Female
Age group	0–29/30–49/50+
Citizenship	Austria/EU except Austria/Non-EU
Highest level of education	Grade school/apprenticeship, vocational school/high- or secondary school, university
Health impairment	Yes/No
Obligations of care (only women)	Yes/No
Occupational group	Production sector/service sector
Regional labor market	Five categories for employment prospects in assigned AMS job center
Prior occupational career	Characterization of variable listed in **Table 2**

Source: Doris Allhutter et al., 'Algorithmic Profiling of Job Seekers in Austria: How Austerity Politics Are Made Effective' (2020) Frontiers in Big Data

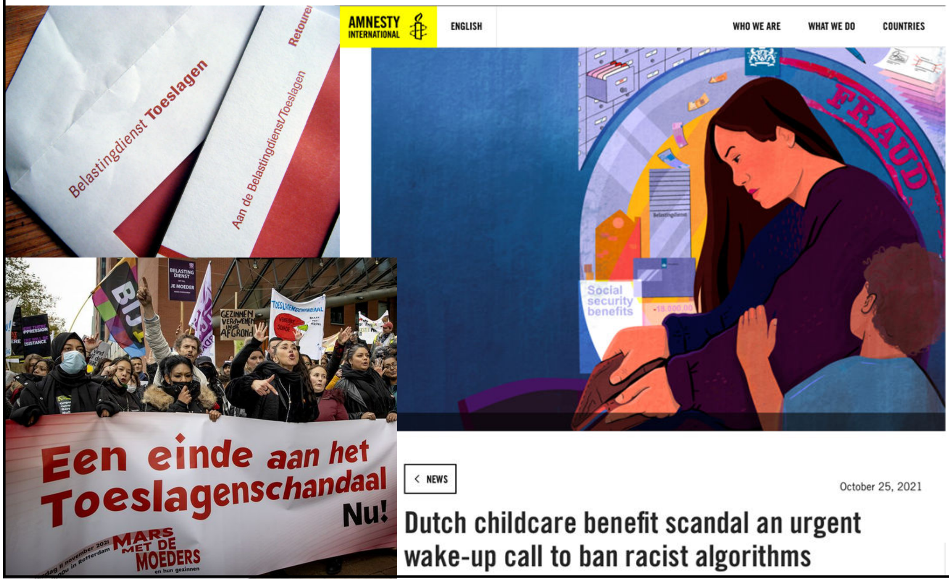
AMS
Arbeitsmarktservice Österreich



Source: https://amsalgorithmus.at/en/

AI'S BROKEN PROMISES

VIGNETTE III: FRAUD DETECTION AND THE WELFARE STATE



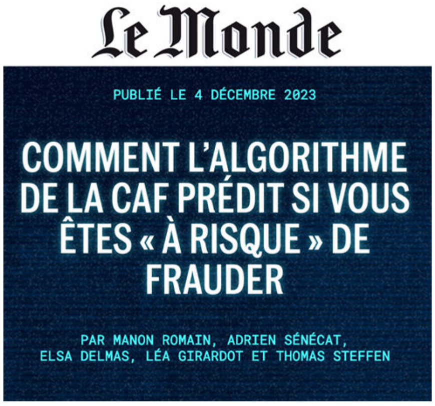
October 25, 2021

Dutch childcare benefit scandal an urgent wake-up call to ban racist algorithms

Source: Amnesty International, Xenophobic machines: Discrimination through unregulated use of algorithms in the Dutch childcare benefits scandal (2021)

AI'S BROKEN PROMISES

VIGNETTE III: FRAUD DETECTION AND THE WELFARE STATE



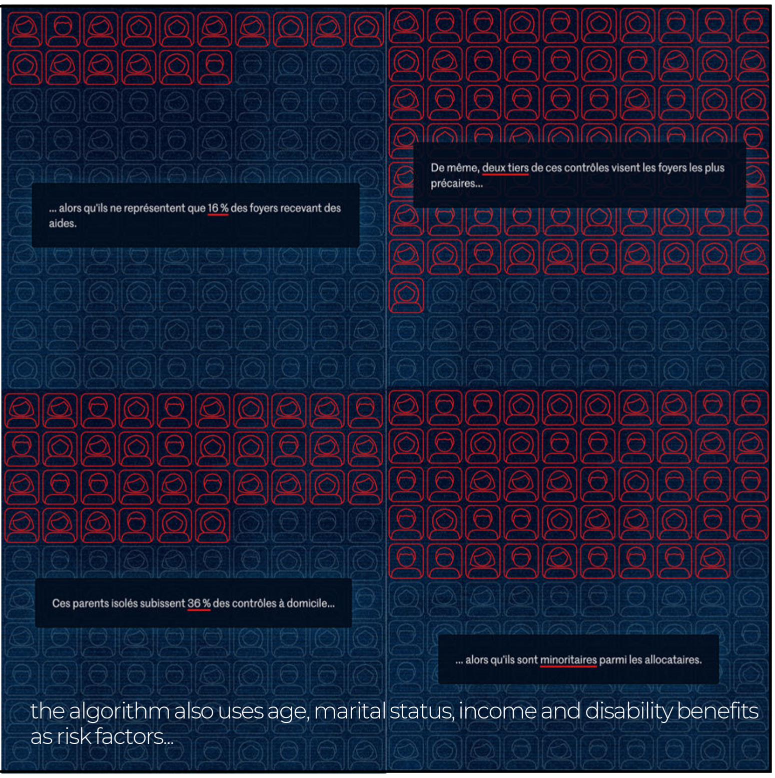
Le Monde

PUBLIÉ LE 4 DÉCEMBRE 2023

COMMENT L'ALGORITHME DE LA CAF PRÉDIT SI VOUS ÊTES « À RISQUE » DE FRAUDER

PAR MANON ROMAIN, ADRIEN SÉNÉCAT, ELSA DELMAS, LÉA GIRARDOT ET THOMAS STEFFEN

In collaboration with **Lighthouse Reports** and **La Quadrature du Net**



… alors qu'ils ne représentent que 16 % des foyers recevant des aides.

De même, deux tiers de ces contrôles visent les foyers les plus précaires…

Ces parents isolés subissent 36 % des contrôles à domicile…

… alors qu'ils sont minoritaires parmi les allocataires.

the algorithm also uses age, marital status, income and disability benefits as risk factors…

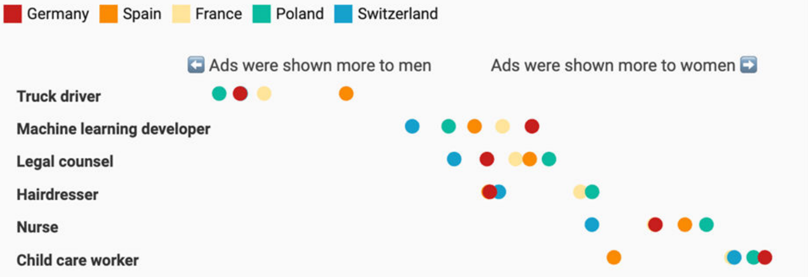
AI'S BROKEN PROMISES

VIGNETTE IV: PROFILING AND AD TARGETING

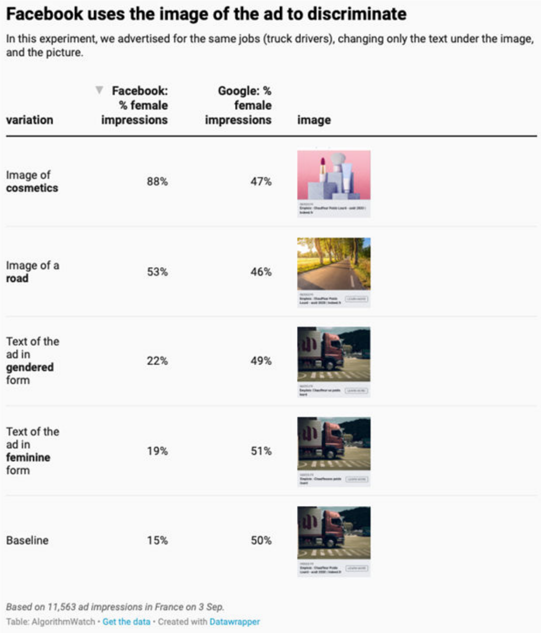
Source: AlgorithmWatch
& Nicolas Kayser-Bril, 'Automated discrimination: Facebook uses gross stereotypes to optimize ad delivery' (2020) AlgorithmWatch available at https://algorithmwatch.org/en/automated-discrimination-facebook-google/

When deciding who to show an ad to, Facebook relies on gross stereotypes

We bought ads for six different job offers in five countries. This is how Facebook optimized the ad impressions, based on gender.



Based on 102,472 ad impressions between 27 Aug and 3 Sep.



Facebook uses the image of the ad to discriminate

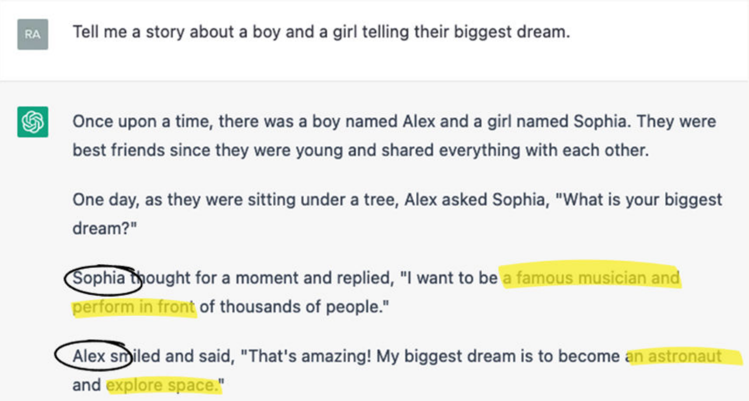
In this experiment, we advertised for the same jobs (truck drivers), changing only the text under the image, and the picture.

variation	▼ Facebook: % female impressions	Google: % female impressions	image
Image of **cosmetics**	88%	47%	
Image of a **road**	53%	46%	
Text of the ad in **gendered** form	22%	49%	
Text of the ad in **feminine** form	19%	51%	
Baseline	15%	50%	

Based on 11,563 ad impressions in France on 3 Sep.

Table: AlgorithmWatch • Get the data • Created with Datawrapper

Open AI's Chat GPT & DALLE-2: algorithmic stereotyping



RA Tell me a story about a boy and a girl telling their biggest dream.

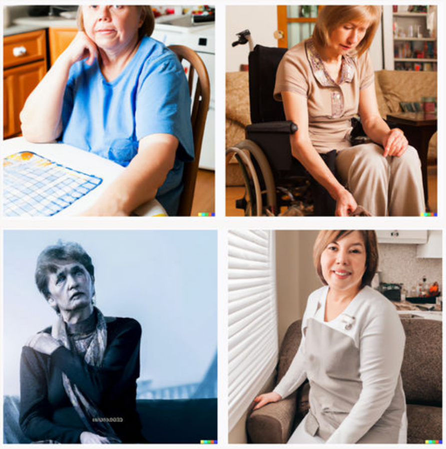
Once upon a time, there was a boy named Alex and a girl named Sophia. They were best friends since they were young and shared everything with each other.

One day, as they were sitting under a tree, Alex asked Sophia, "What is your biggest dream?"

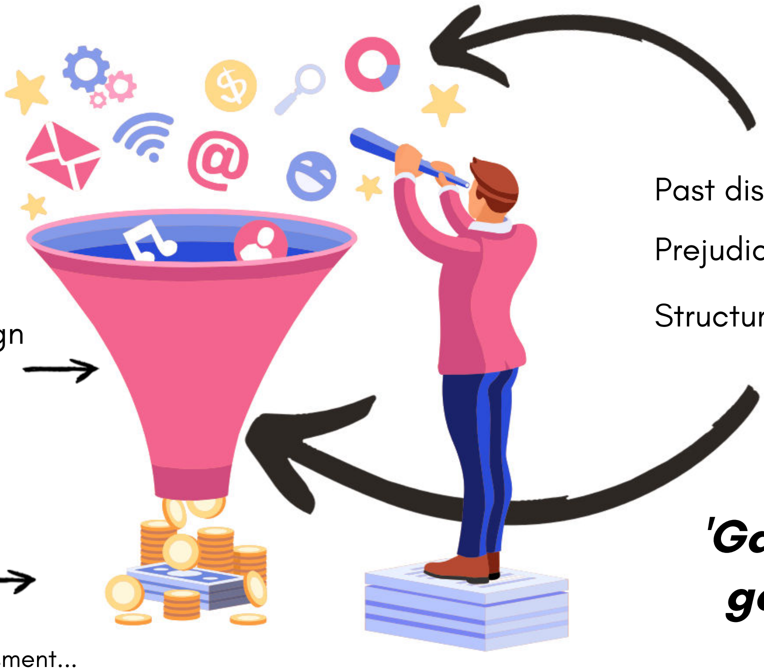
Sophia thought for a moment and replied, "I want to be a famous musician and perform in front of thousands of people."

Alex smiled and said, "That's amazing! My biggest dream is to become an astronaut and explore space."

Chat GPT
Experiment performed on 28 March 2023
Inspired by Ivana Bartoletti



DALLE-2
Experiment performed in December 2022 (first outcome)
Prompt used: 'a human caregiver sits passively at home'



Data →

Problem design and model →

Output: →
recommendation, prediction, assessment...

Past discrimination

Prejudices, stereotypes

Structural inequalities

'Garbage in, garbage out'

Source: DALL-E 2

2. What is unlawful discrimination under EU law?

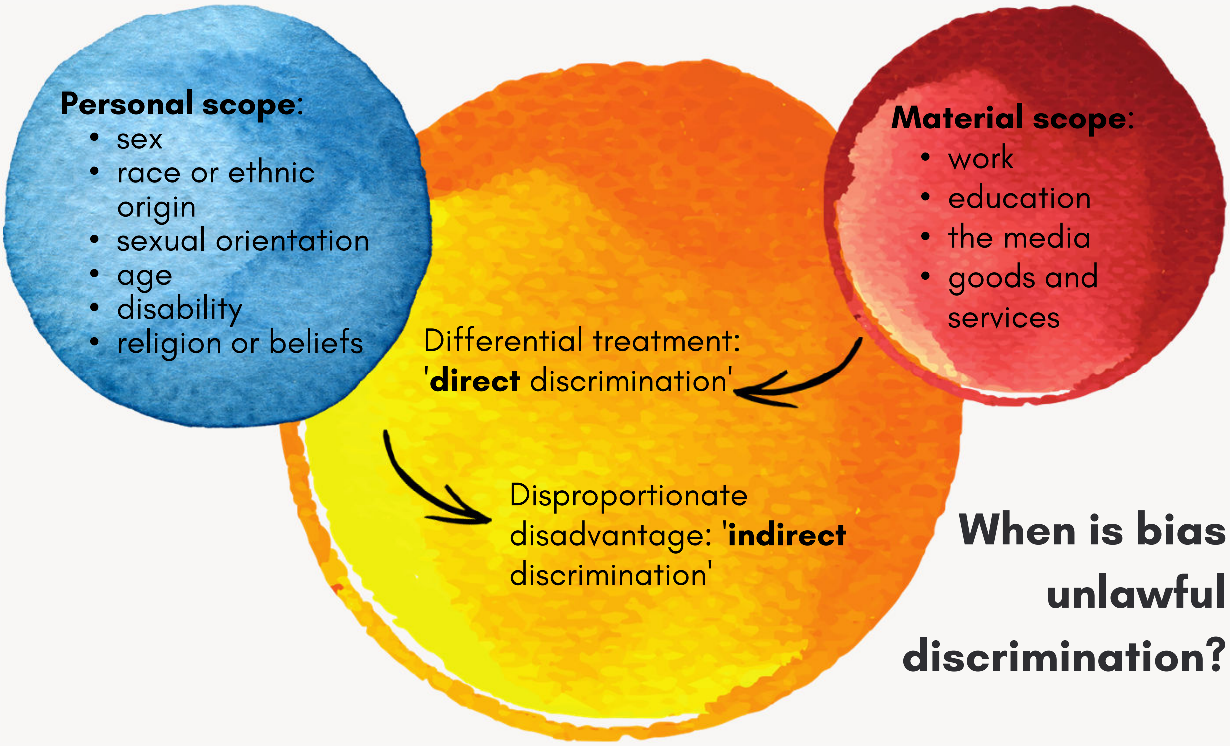
A regulatory puzzle: EU anti-discrimination law

- **Art 19 TFEU**: sex, race or ethnic origin, religion or beliefs, disability, age, sexual orientation
- **Art 157 TFEU**: equality between men and women at work and in pay
- **Art 21 Charter of Fundamental Rights**: non-exhaustive list of protected criteria including
- **Art 23 Charter**: equality between men and women

Minimum requirements

- **Directive 2000/43**: race or ethnic origin (employment, goods and services, education)
- **Directive 2000/78**: age, religion or belief, disability, sexual orientation (employment)
- **Directive 2004/113**: sex (goods and services)
- **Directive 2006/54**: sex (employment)

When is bias unlawful discrimination?



Personal scope:
- sex
- race or ethnic origin
- sexual orientation
- age
- disability
- religion or beliefs

Material scope:
- work
- education
- the media
- goods and services

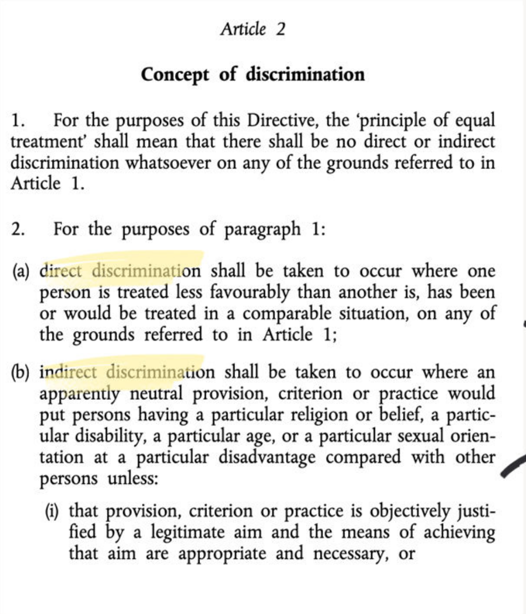
Differential treatment: '**direct** discrimination'

Disproportionate disadvantage: '**indirect** discrimination'

Source: DALL-E 2

3. How to redress algorithmic discrimination?

Legal consequences of the direct/indirect qualification



Article 2

Concept of discrimination

1. For the purposes of this Directive, the 'principle of equal treatment' shall mean that there shall be no direct or indirect discrimination whatsoever on any of the grounds referred to in Article 1.

2. For the purposes of paragraph 1:

(a) direct discrimination shall be taken to occur where one person is treated less favourably than another is, has been or would be treated in a comparable situation, on any of the grounds referred to in Article 1;

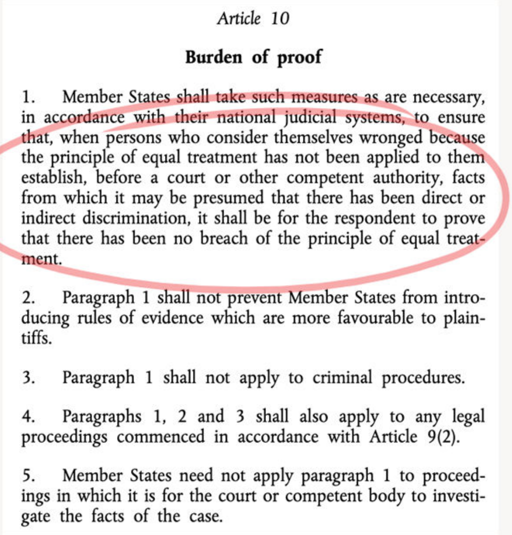
(b) indirect discrimination shall be taken to occur where an apparently neutral provision, criterion or practice would put persons having a particular religion or belief, a particular disability, a particular age, or a particular sexual orientation at a particular disadvantage compared with other persons unless:

(i) that provision, criterion or practice is objectively justified by a legitimate aim and the means of achieving that aim are appropriate and necessary, or

No justification in principle, save for exceptions, e.g. genuine and determining occupational requirement

Open-ended justification regime: proportionality test (pure economic justifications excluded in principle)

Shift of the burden of proof



Article 10

Burden of proof

1. Member States shall take such measures as are necessary, in accordance with their national judicial systems, to ensure that, when persons who consider themselves wronged because the principle of equal treatment has not been applied to them establish, before a court or other competent authority, facts from which it may be presumed that there has been direct or indirect discrimination, it shall be for the respondent to prove that there has been no breach of the principle of equal treatment.

2. Paragraph 1 shall not prevent Member States from introducing rules of evidence which are more favourable to plaintiffs.

3. Paragraph 1 shall not apply to criminal procedures.

4. Paragraphs 1, 2 and 3 shall also apply to any legal proceedings commenced in accordance with Article 9(2).

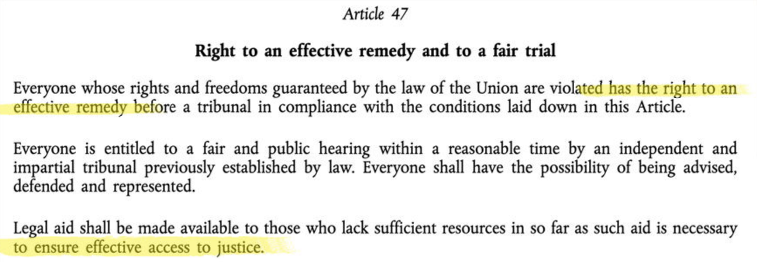
5. Member States need not apply paragraph 1 to proceedings in which it is for the court or competent body to investigate the facts of the case.

- Applicants are required to establish a ***prima facie* case** of discrimination

- NB: **no proof of intent** required for a finding of direct discrimination
- **C-415/10 *Meister***: no right to information about recruitment process
- But **C-109/88 *Danfoss*** [13]: 'where a system of individual pay supplements which is **completely lacking in transparency** is at issue, female employees can establish differences only so far as average pay is concerned. **They would be deprived of any effective means of enforcing the principle of equal pay** before the national courts **if the effect of adducing such evidence was not to impose upon the employer the burden of proving** that his practice in the matter of wages is not in fact discriminatory'

Effective remedies

EU Charter of Fundamental Rights



Article 47

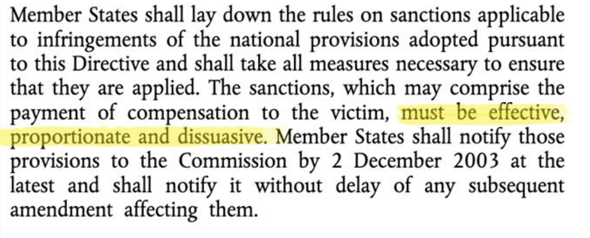
Right to an effective remedy and to a fair trial

Everyone whose rights and freedoms guaranteed by the law of the Union are violated has the right to an effective remedy before a tribunal in compliance with the conditions laid down in this Article.

Everyone is entitled to a fair and public hearing within a reasonable time by an independent and impartial tribunal previously established by law. Everyone shall have the possibility of being advised, defended and represented.

Legal aid shall be made available to those who lack sufficient resources in so far as such aid is necessary to ensure effective access to justice.

Directive 2000/78/EC



Article 17

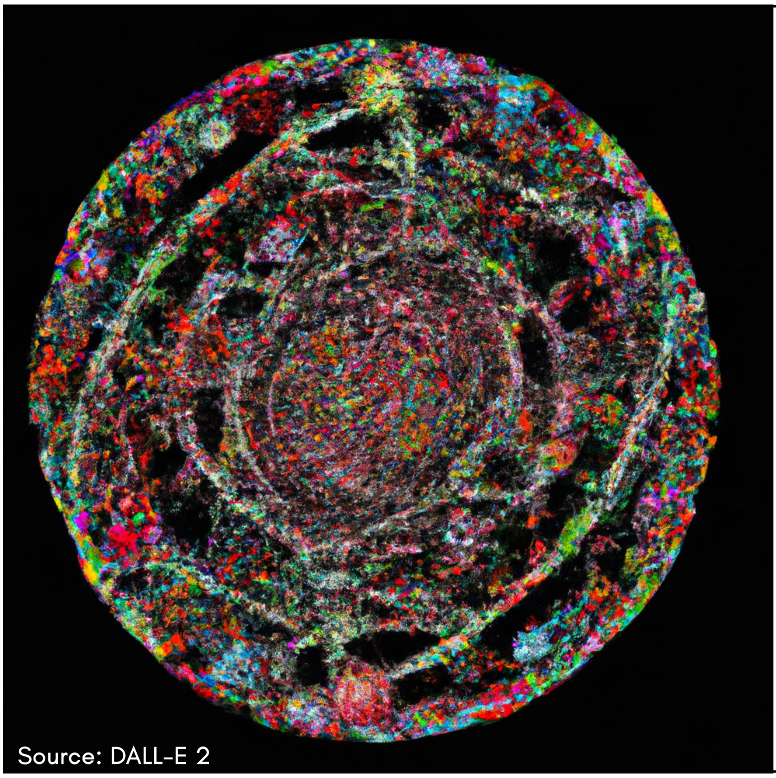
Sanctions

Member States shall lay down the rules on sanctions applicable to infringements of the national provisions adopted pursuant to this Directive and shall take all measures necessary to ensure that they are applied. The sanctions, which may comprise the payment of compensation to the victim, must be effective, proportionate and dissuasive. Member States shall notify those provisions to the Commission by 2 December 2003 at the latest and shall notify it without delay of any subsequent amendment affecting them.

What remedies?

On sanctions – C-54/07 *Firma Feryn* [39]:
'If it appears appropriate to the situation at issue in the main proceedings, those sanctions may, where necessary, include **a finding of discrimination** by the court or the competent administrative authority **in conjunction with an adequate level of publicity, the cost of which is to be borne by the defendant**. They may also take the form of a **prohibitory injunction, in accordance with the rules of national law, ordering the employer to cease the discriminatory practice**, and, where appropriate, **a fine**. They may, moreover, take the form of the **award of damages to the body bringing the proceedings**.'

No levelling down – C-193/17 *Cresco* [79]:
In the fourth place, it should be noted that, according to settled case-law of the Court, where discrimination contrary to EU law has been established, **as long as measures reinstating equal treatment have not been adopted, observance of the principle of equality can be ensured only by granting to persons within the disadvantaged category the same advantages as those enjoyed by persons within the favoured category**. Disadvantaged persons must therefore be placed in the same position as persons enjoying the advantage concerned (judgment of 9 March 2017, Milkova, C-406/15, EU:C:2017:198, paragraph 66 and the case-law cited).


Source: DALL-E 2

4. The interaction between EU anti-discrimination law, data protection and AI regulation

WHAT DOES THE LAW SAY ABOUT ALGORITHMIC BIAS?

Overview

FUNDAMENTAL RIGHTS

EU Charter, EU Treaties, European Convention on Human Rights

MINIMUM REQUIREMENTS

EU anti-discrimination law

AI SECTORAL REGULATION

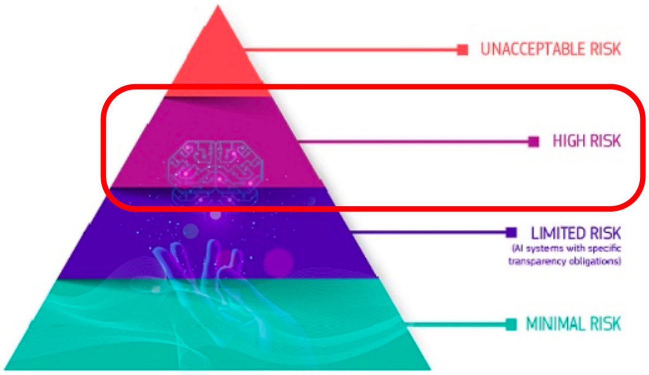
EU AI Act, Council of Europe Framework Convention on AI, Proposal for an AI Liability Directive

OTHER RELEVANT LEGAL PROVISIONS

Consumer protection, Data protection law....

Proposal for a Regulation laying down harmonised rules on artificial intelligence (EU AI Act)

- Risk-based approach
- Liability: AI Liability Directive



Source: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

Zooming on the EU AI Act (common approach Nov 2022)

- Explanatory memorandum & recital 28: when classifying an AI system as high-risk, it is of particular relevance to consider '[t]he extent of the **adverse impact caused by the AI system** on the fundamental rights protected by the Charter' including '**non-discrimination**' (Art. 21 EUCFR) and '**equality between women and men**' (Art. 23 EUCFR)

- Recitals 35, 36 and 37 warn that AI systems used in core sectors such as education, employment and essential services are liable to '**violate [...] the right not to be discriminated against**' and '**perpetuate historical patterns of discrimination**'

- Recital 44 explicitly refers to non-discrimination law when stressing the **importance of high-quality data requirements to ensure that a high-risk AI system 'does not become the source of discrimination prohibited by Union law'**

The EU AI Act: requirements for data

Article 10
Data and data governance

1. **High-risk AI systems which make use of techniques involving the training of models with data shall be developed on the basis of training, validation and testing data sets that meet the quality criteria referred to in paragraphs 2 to 5.**
2. Training, validation and testing data sets shall be subject to **appropriate data governance and management practices.** Those practices shall concern in particular,
(a) the relevant **design choices;**
(b) **data collection** processes;
(c) relevant **data preparation** processing operations, such as **annotation, labelling, cleaning, enrichment and aggregation;**
(d) the **formulation of relevant assumptions**, notably with respect to **the information that the data are supposed to measure and represent;**
(e) a p**rior assessment of the availability, quantity and suitability of the data sets** that are needed;
(f) examination in view of possible biases that are likely to affect health and safety of natural persons or lead to discrimination prohibited by Union law;
(g) the **identification of any possible data gaps or shortcomings**, and how those gaps and shortcomings can be addressed.

The GDPR and 'sensitive data'

GDPR Article 9
Processing of special categories of personal data

1. Processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation ***shall be prohibited****.*

→ **Does data protection law allow using sensitive data to detect and correct algorithmic bias?**

The EU AI Act and the GDPR

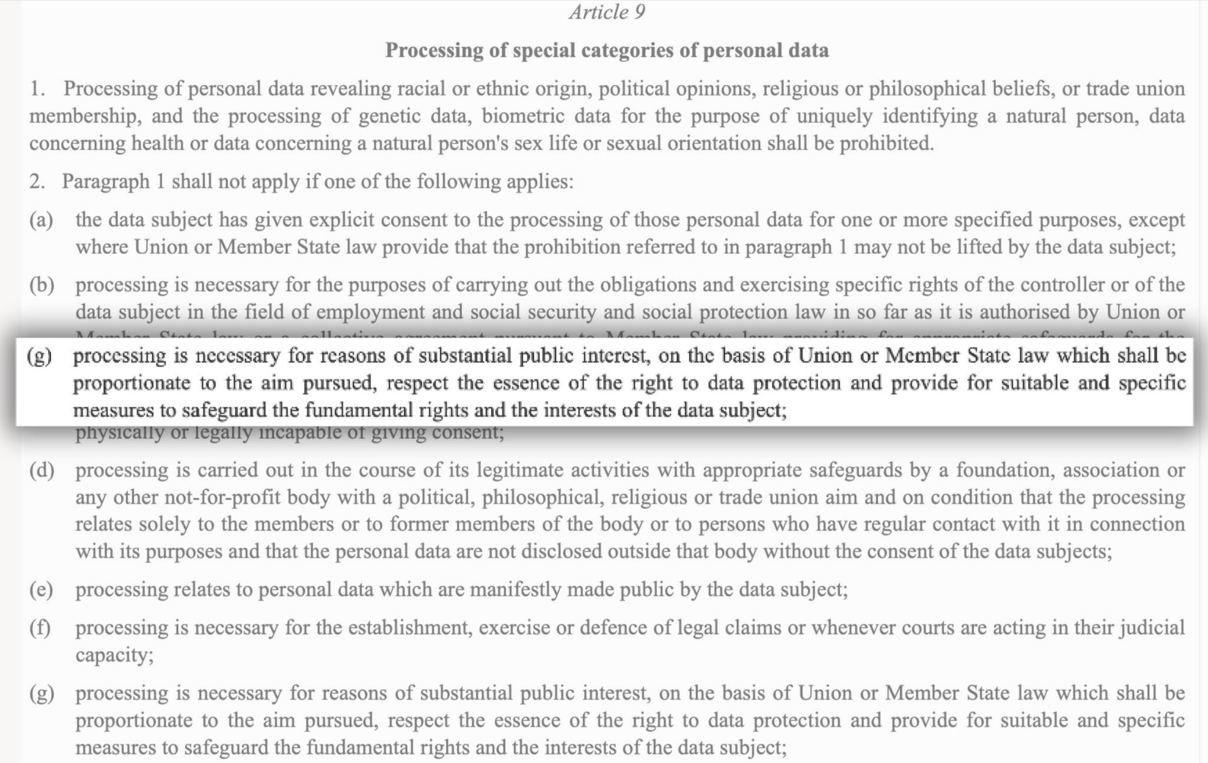
Article 10
Data and data governance

3. **Training, validation and testing data sets shall be relevant, representative, and to the best extent possible, free of errors and complete.** They shall have the appropriate statistical properties, including, where applicable, as regards the persons or groups of persons on which the high-risk AI system is intended to be used. These characteristics of the data sets may be met at the level of individual data sets or a combination thereof.

4. Training, validation and testing data sets shall **take into account, to the extent required by the intended purpose, the characteristics or elements that are particular to the specific geographical, behavioural or functional setting within which the high-risk AI system is intended to be used.**

5. To the extent that it is strictly necessary for the purposes of ensuring bias monitoring, detection and correction in relation to the high-risk AI systems, the providers of such systems may process special categories of personal data referred to in Article 9(1) of [the GDPR], Article 10 of Directive (EU) 2016/680 [on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data] and Article 10(1) of Regulation (EU) 2018/1725 [on the protection of natural persons with regard to the processing of personal data by the Union institutions, bodies, offices and agencies and on the free movement of such data], **subject to appropriate safeguards for the fundamental rights and freedoms of natural persons,** including technical limitations on the re-use and use of state-of-the-art security and privacy-preserving measures, such as pseudonymisation, or encryption where anonymisation may significantly affect the purpose pursued.

The GDPR



Article 9

Processing of special categories of personal data

1. Processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation shall be prohibited.

2. Paragraph 1 shall not apply if one of the following applies:

(a) the data subject has given explicit consent to the processing of those personal data for one or more specified purposes, except where Union or Member State law provide that the prohibition referred to in paragraph 1 may not be lifted by the data subject;

(b) processing is necessary for the purposes of carrying out the obligations and exercising specific rights of the controller or of the data subject in the field of employment and social security and social protection law in so far as it is authorised by Union or

(g) processing is necessary for reasons of substantial public interest, on the basis of Union or Member State law which shall be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject;

physically or legally incapable of giving consent;

(d) processing is carried out in the course of its legitimate activities with appropriate safeguards by a foundation, association or any other not-for-profit body with a political, philosophical, religious or trade union aim and on condition that the processing relates solely to the members or to former members of the body or to persons who have regular contact with it in connection with its purposes and that the personal data are not disclosed outside that body without the consent of the data subjects;

(e) processing relates to personal data which are manifestly made public by the data subject;

(f) processing is necessary for the establishment, exercise or defence of legal claims or whenever courts are acting in their judicial capacity;

(g) processing is necessary for reasons of substantial public interest, on the basis of Union or Member State law which shall be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject;

Source: DALL-E 2

5. Shortcomings and remaining issues

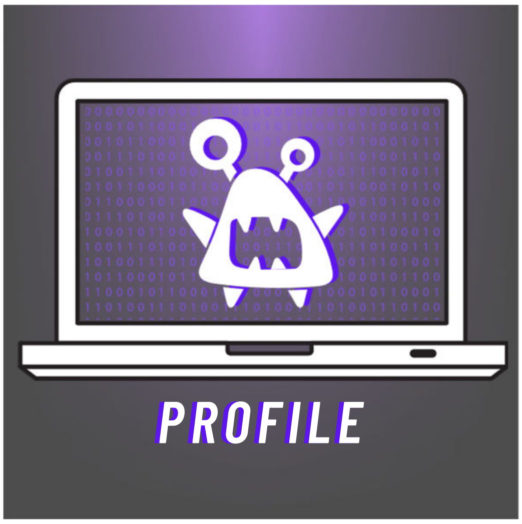
Redressing algorithmic discrimination: detection and enforcement

- **Problem 1**: heuristics of discrimination law, the comparison test, disrupted
  - Online personalisation: end users 'trapped' inside own reality, suspicion of discrimination not raised
  - *Prima facie* evidence difficult to bring
- **Problem 2**: causality difficult to prove
  - 'Black box' algorithms: no access to meaningful information (no transparency, no explainability, opacity of proprietary algorithms)
  - Respondents may have difficulty rebutting a *prima facie* case of discrimination
  - blurred direct/indirect discrimination distinction
- **Problem 3**: the scope of discrimination law
  - patchy material scope
  - limited personal scope vs emergent discrimination
  - intersectional discrimination
- **Problem 4**: liability of providers vs users (complex human-machine interactions, automation biases)

EU AIA, AILD and ADL: do they fit together?

- **EU AIA:**
  - High-risk systems vs **systemic discrimination**
  - No actionable **individual rights**
  - **Access to justice**?
    - Information on use of AI
    - Role for public supervision (equality bodies)?

- **AI Liability Directive**: facilitating enforcement of ADL rules?
  - **Disclosure of evidence**: access to evidence in the context of information asymmetries
  - But can that help adduce *prima facie* evidence? no 'blanket requests' + 'in support of that request, the potential claimant must present facts and evidence sufficient to support the plausibility of a claim for damages'
  - Art. 3(5): **presumption of non-compliance**
  - **Rebuttable presumption of a causal link in the case of fault**: fit with no-fault liability regime?
  - At the level of **prevention**: a role for negligence and *ex ante* impact assessments? Inclusion through the concept of 'instruction to discriminate' and inspiration from vicarious liability?



THANK YOU!

equalgo.wordpress.com

@RaphaeleXenidis

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 898937.