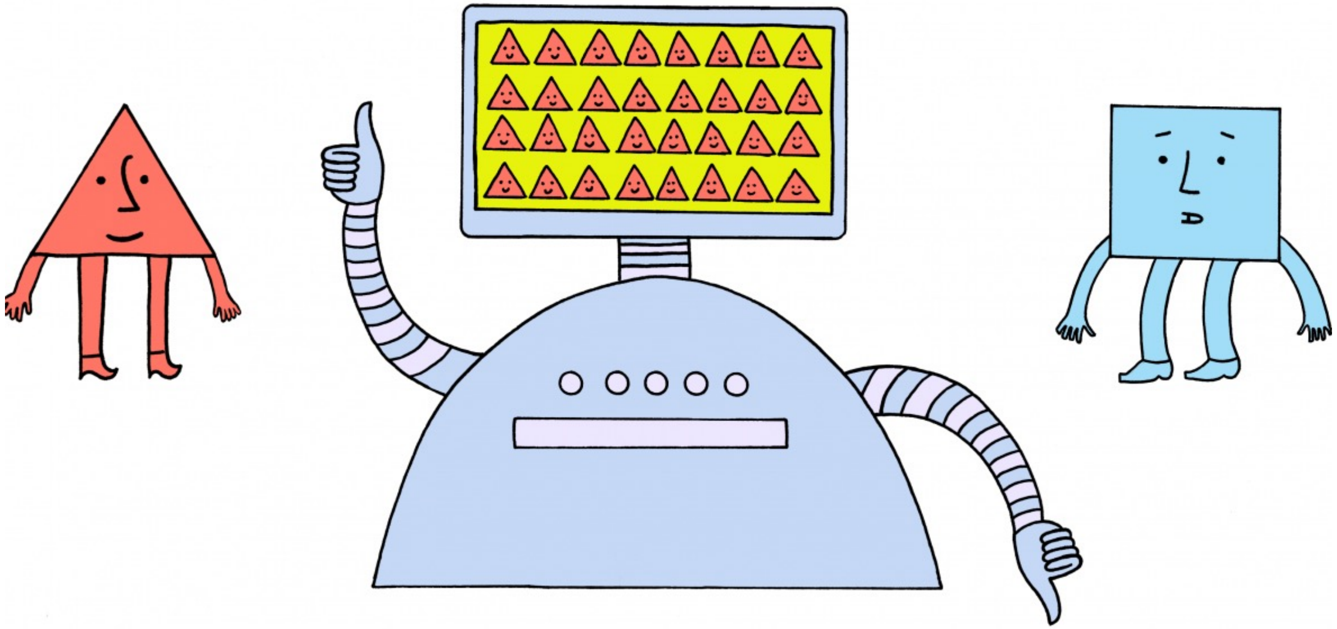


Algorithmic bias and algorithmic governance

CT5142 Ethics and AI

Content

- Fairness and the ethics of algorithms
- The problem of bias
- Algorithmic governance and the problem of power(lessness)
- Algorithmic accountability

Algorithmic fairness and the ethics of algorithms

Justice as ethical principle

⚖ Justice addresses the balance of vulnerabilities and needs between people

Equality and non-discrimination

₿ Fair distribution of benefits and burdens

Who are the people who benefit, who carries a burden? Any discrepancies?

Are any groups of people and their concerns excluded from benefits?

STOA (2019): A governance system of algorithmic accountability and transparency

The problem:

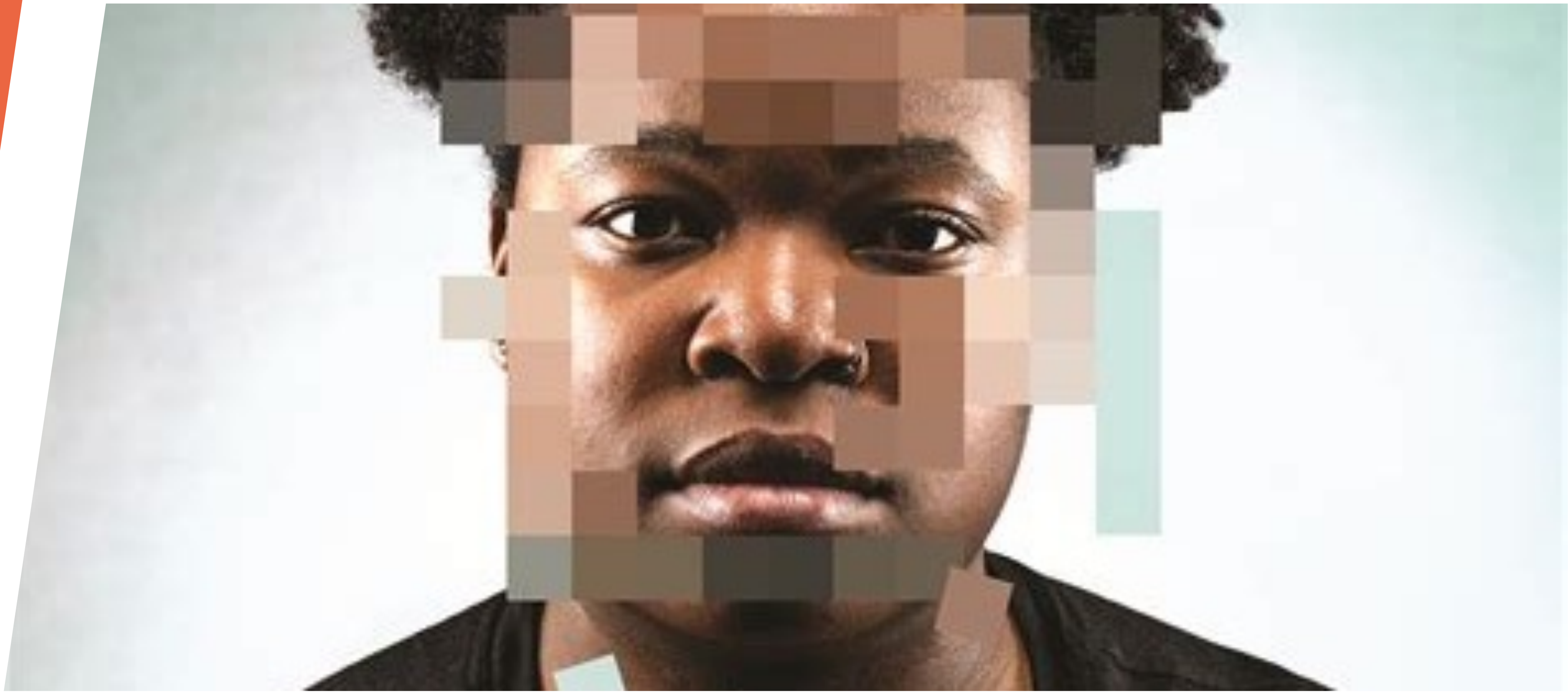
- “Algorithmic systems are increasingly being used as part of decision-making processes with potentially significant consequences for individuals, organisations and societies as a whole. Because the ways in which these systems reach their 'conclusions' may reflect or amplify existing biases, or may not offer explanations that satisfy our accustomed social and judicial expectations, there is growing concern that the traditional frameworks for implementing transparency and accountability may not suffice as mechanisms of governance.” (p.1)

The goal:

- Algorithmic fairness, understood as “fairness in the context of algorithmic implementations appears as a balance between the mutual interests, needs and values of different stakeholders affected by the algorithmic decision” (p.10)

Mittelstadt et al. (2016) The Ethics of Algorithms : core problems

Inconclusive evidence	Data might not be sufficient to draw well-grounded conclusions even if there are patterns indicated
Inscrutable evidence	Often “logical”/causal connection is not evident
Misguided evidence	“garbage in - garbage out”
Unfair outcomes	Unfair discrimination actions may result even from high quality data
Transformative effects	May lead us to reconceptualise the world according to results
Traceability	Often not clear where mistake/harm originates



The problem of algorithmic bias

Bias

Bias:

Systematic error, based on faulty selection

A general tendency to judge in a particular way, based on prejudice, or giving undue weight to some considerations over others

Can lead to unjust discrimination

Unconscious/implicit bias:

Bias that is not recognised but comes out in action

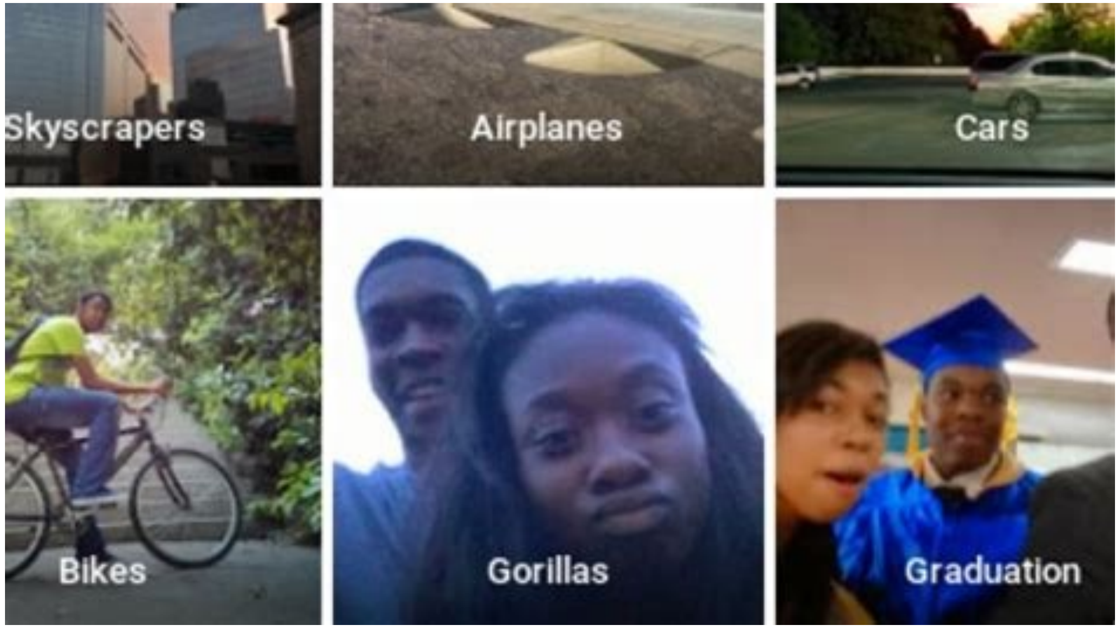
Unconscious biases psychologically common, and often resistant to correction

Algorithmic bias:

Bias that is encoded in algorithms (see Nissenbaum & Friedman on bias and embedded values, WK1)

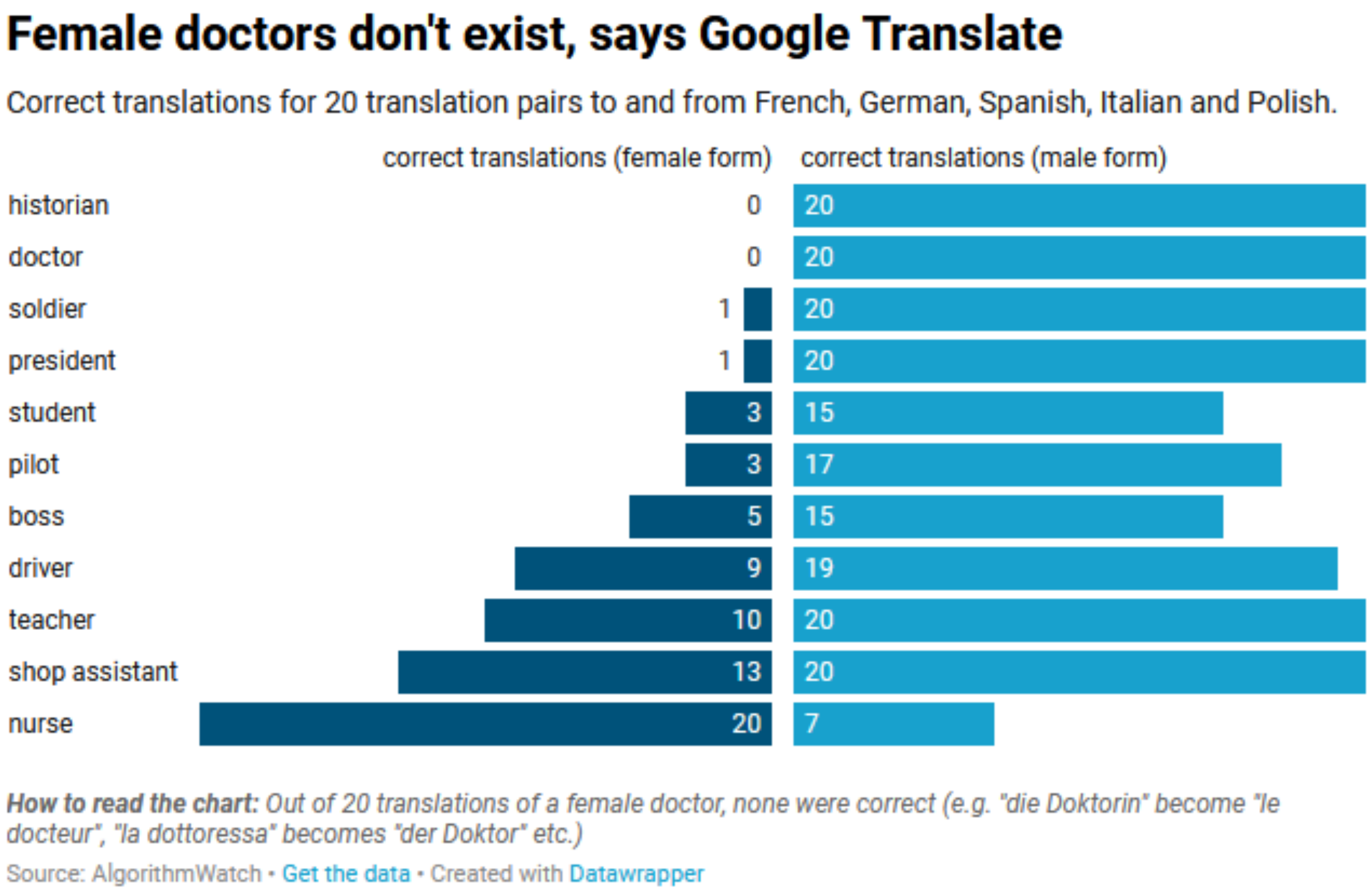
Can derive e.g. from incomplete data, biased training datasets, technology designers with implicit biases

Racial bias



- the racist soap dispenser
- Low accuracy of face recognition technologies for non-white, non-male face
- Google‘s classification of black faces as „gorillas“

Systematic gender mistranslations on Google Translate



Female doctors don't exist, says Google Translate

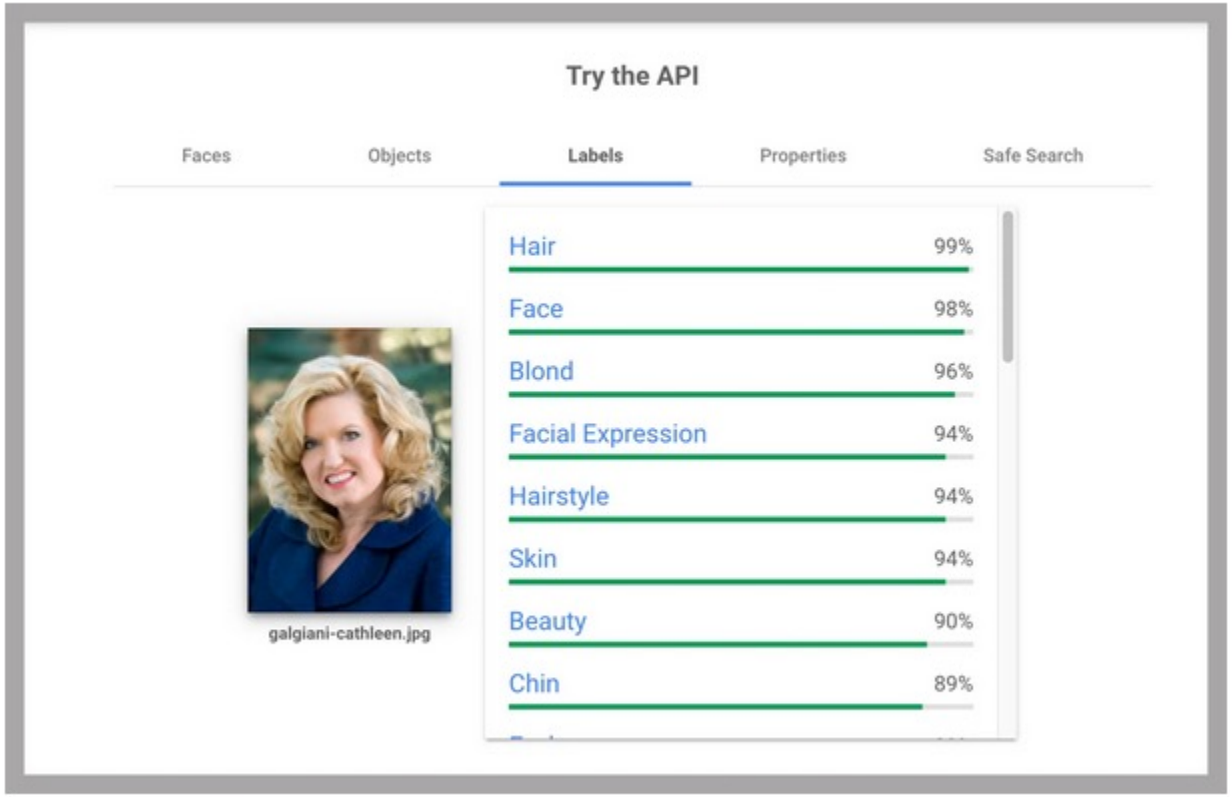
Correct translations for 20 translation pairs to and from French, German, Spanish, Italian and Polish.

How to read the chart: *Out of 20 translations of a female doctor, none were correct (e.g. "die Doktorin" become "le docteur", "la dottoressa" becomes "der Doktor" etc.)*

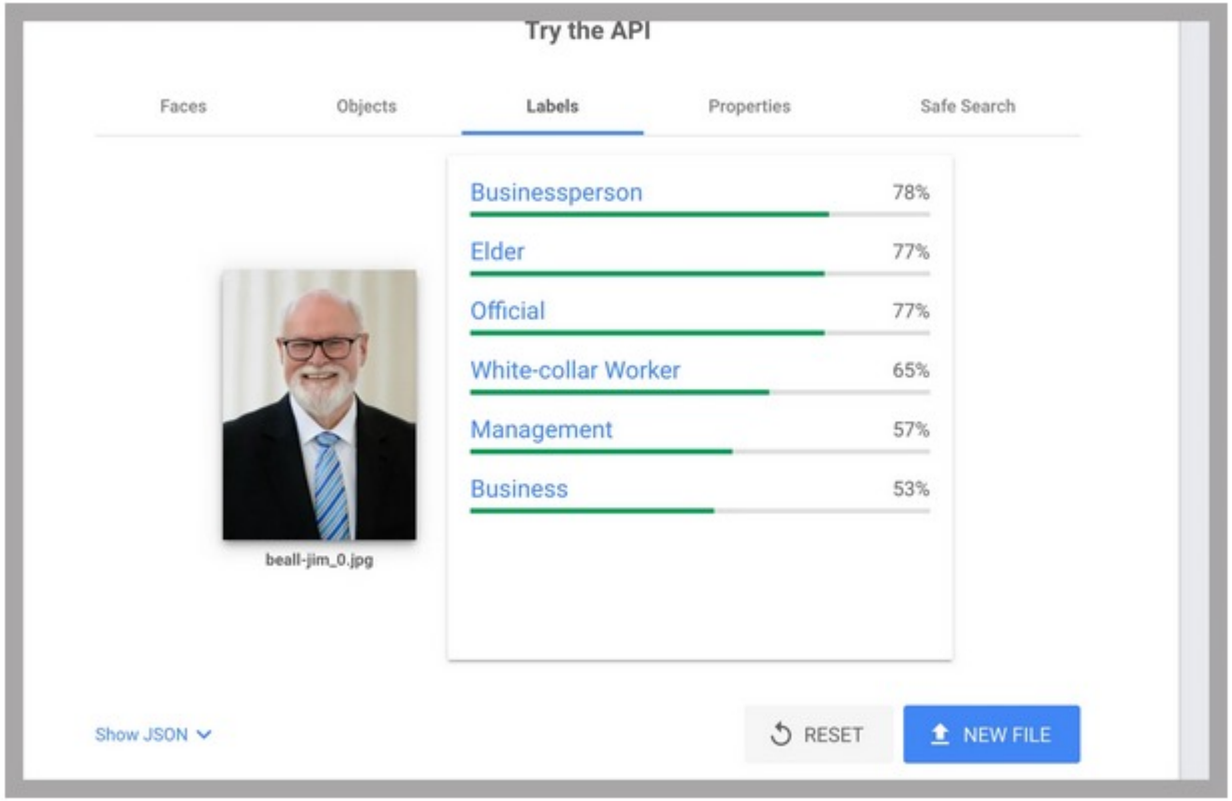
Source: AlgorithmWatch • Get the data • Created with Datawrapper

https://algorithmwatch.org/en/story/google-translate-gender-bias/ (September 2020)

AI image content labelling: women‘s appearance, men‘s profession



Hair	99%
Face	98%
Blond	96%
Facial Expression	94%
Hairstyle	94%
Skin	94%
Beauty	90%
Chin	89%

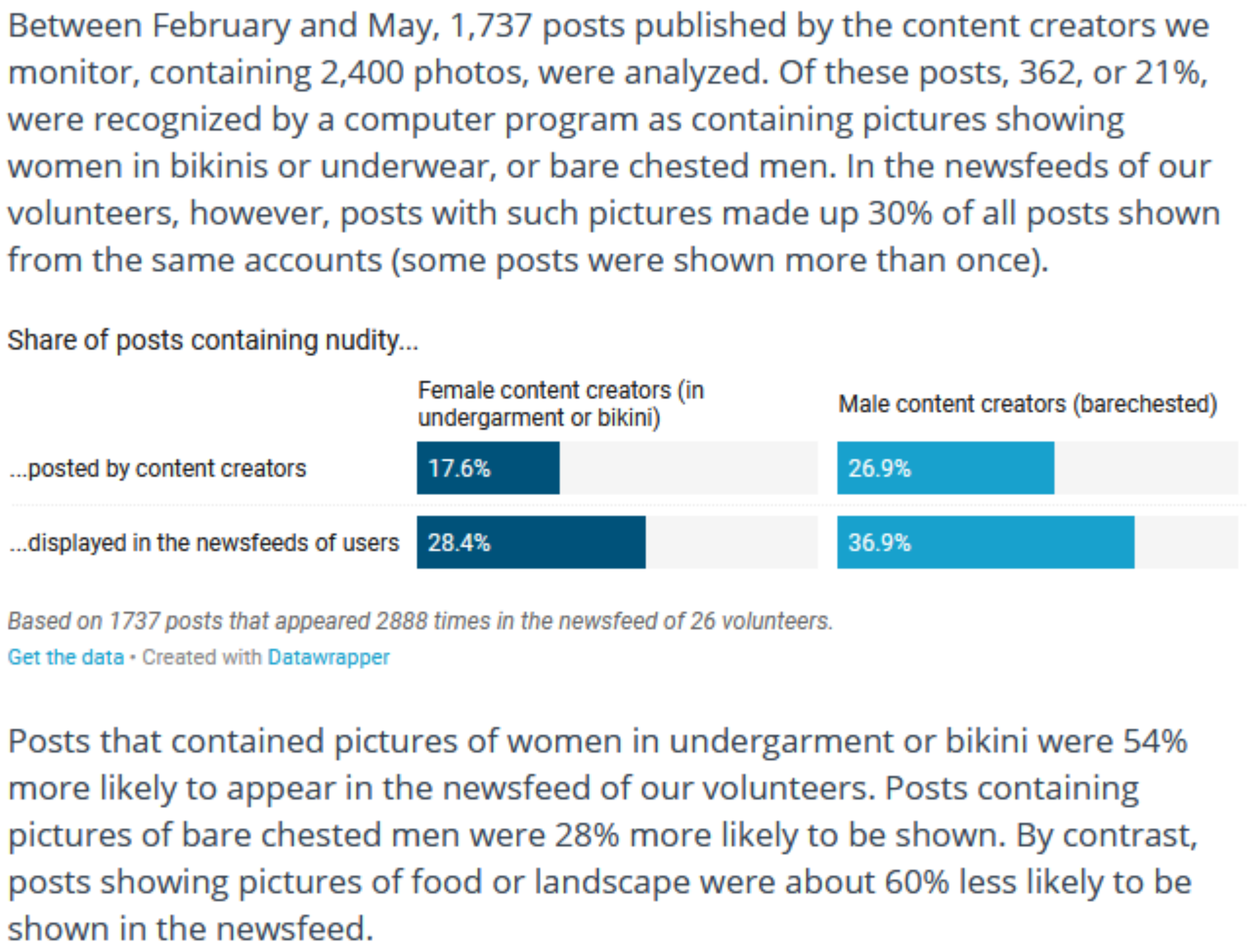


Businessperson	78%
Elder	77%
Official	77%
White-collar Worker	65%
Management	57%
Business	53%

WIRED STAFF VIA GOOGLE

Tom Simonite (2020). When AI Sees a Man, It Thinks 'Official.' A Woman? 'Smile’, Wired 19/11/20
https://www.wired.com/story/ai-sees-man-thinks-official-woman-smile/
Referring to Schwemmer et al. (2020) Diagnosis Gender Bias in Image Recognition Systems, *Socius*,
https://journals.sagepub.com/doi/10.1177/2378023120967171

Instagram image display biased towards female bare skin (Algorithm Watch 2020)



Between February and May, 1,737 posts published by the content creators we monitor, containing 2,400 photos, were analyzed. Of these posts, 362, or 21%, were recognized by a computer program as containing pictures showing women in bikinis or underwear, or bare chested men. In the newsfeeds of our volunteers, however, posts with such pictures made up 30% of all posts shown from the same accounts (some posts were shown more than once).

Share of posts containing nudity...

	Female content creators (in undergarment or bikini)	Male content creators (barechested)
...posted by content creators	17.6%	26.9%
...displayed in the newsfeeds of users	28.4%	36.9%

Based on 1737 posts that appeared 2888 times in the newsfeed of 26 volunteers.

Get the data • Created with Datawrapper

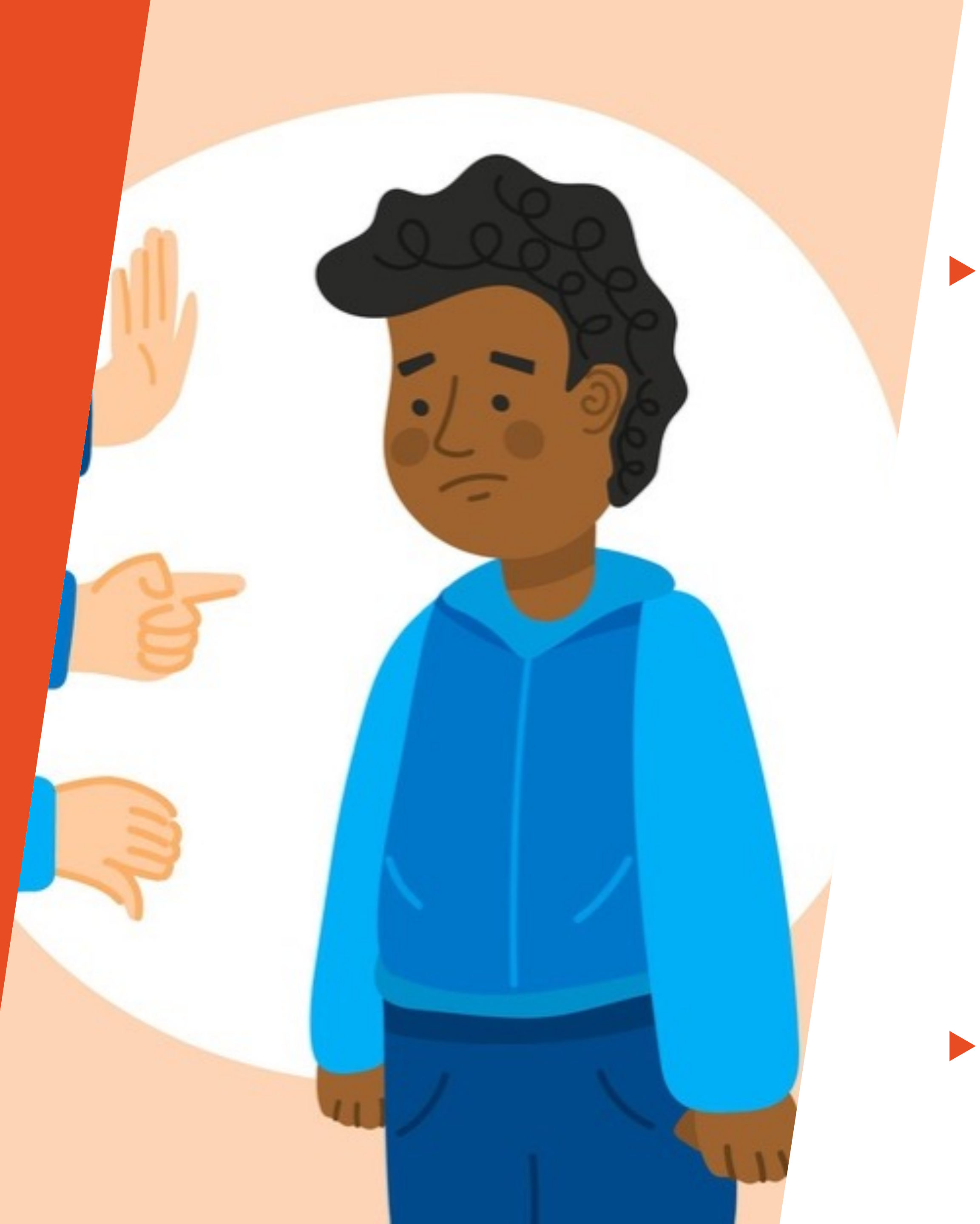
Posts that contained pictures of women in undergarment or bikini were 54% more likely to appear in the newsfeed of our volunteers. Posts containing pictures of bare chested men were 28% more likely to be shown. By contrast, posts showing pictures of food or landscape were about 60% less likely to be shown in the newsfeed.

Example: Amazon‘s biased algorithmic hiring tool

- Use of computerized hiring tool, using artificial intelligence to score job applicants and predict which of the top candidates would be best for the job.
- It turned out that the software was discriminating strongly against women, e.g. negative weighting assigned to graduating from all-women's colleges or just including the word "women's" on their CV
- problem due to the data used for training the tool: based on CVs received by company in prior decade, plus information on whether applicants were hired.
- The vast majority of training CVs were from male applicants, and the majority of hires were men. Drawing conclusions from this data the system assumed that men were generally preferable to women.

Past data and future change

- Algorithms use data representing past practices for impacting the future
- Perpetuating the past by inscribing it into the future
- Especially problematic when applied to social practices that are subject to change or seen as in need of change

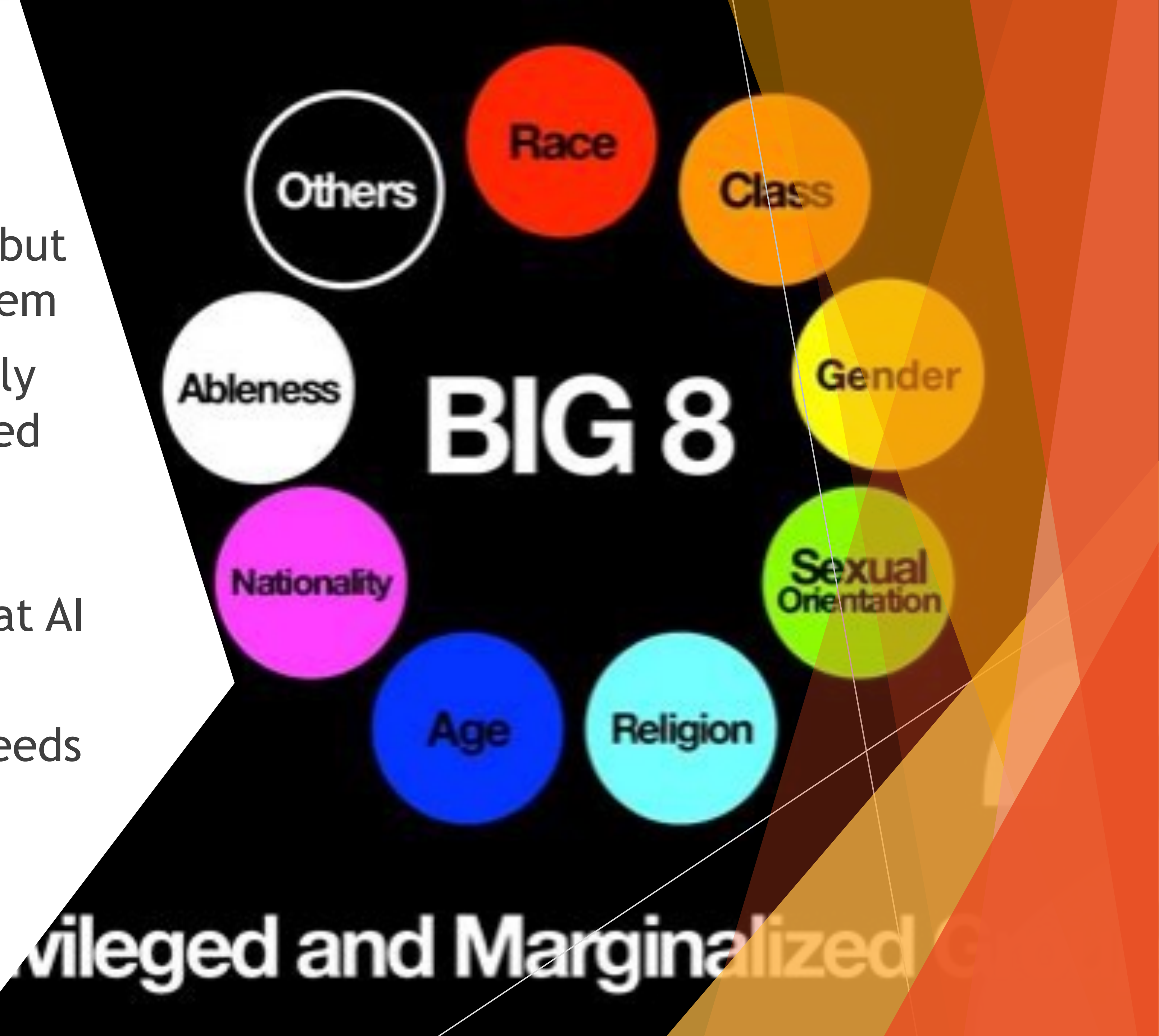


Big data: A tool for inclusion or exclusion? (FTC 2016)

- Algorithmic analysis promises innovations that might allow identification of biases and reduce discrimination and promote fairer allocation of opportunity, e.g.
  - Expansion of access to credit in low-income communities
  - Removal of subconscious bias from hiring decisions and classrooms
  - Provision of extra resources to at-risk students (based on learning analytics)
- However, it might also perpetuate or exacerbate existing inequalities

What to look for?

- “Fairness” not a technical, but an ethical or political problem
- Useful heuristic: traditionally marginalised or discriminated groups
- But also context-specific concerns (depending on what AI system does)
- i.e. not always obvious & needs reflection



Equal opportunity by design?

- Discrimination as significant risk, e.g. with regard to gender, race, national origin, sexual orientation, disability, or socio-economic status (categories covered by Equality legislation in IE)
- Algorithmic decision-making in public and private sectors as potentially risky with regard to citizens‘ rights
- Existing legal instruments not always clearly covered by algorithmic practices (variety of potential legal instruments, from privacy to human rights)
- Potential solution to demand „equal opportunity by design“, similar to privacy by design

Algorithmic governance and the problem of power

Algorithmic governance

- Exerting social power through the medium of algorithms
- Algorithms used as tools of dominance by political actors or administrative system
- Algorithms also shape constructions of social reality, and structure social space and positioning of citizens in novel ways

Power and the asymmetry of algorithmic assessment

- Power as ability to control and influence actions and circumstances affecting oneself and one‘s opportunities vis-a-vis others
- Power closely related to social position and professional roles
- Algorithmic systems are employed within existing social power structures
- They are not just a result and expression of these structures, but can also perpetuate power dynamics
  - E.g. cycle of surveillance: when A gets surveilled by B, B exercises power but also perpetuates power over A due to knowing additional things about A that might lead to more reasons for suspicion and surveillance (whereas A generally never learns things about the surveillance process or B as social actor that would help to correct faulty processes and redress the power imbalance)

Automating inequality (Eubanks 2018): marginalisation through algorithms

- “the digital poorhouse”: algorithmic poverty management with risk of massive scaling up of algorithmic decision-making without sufficient safeguards
- concern that algorithmically supported decision-making affects the most marginalised and vulnerable most strongly
- Long tradition of surveillance and assessment of whether poor are deserving of assistance or sufficiently collaborative
- Examples: automated risk assessment for child protection, welfare eligibility, surveillance of welfare spending

Powerlessness of humans vis-a-vis algorithmic systems (STOA 2019)

„Unless appropriate governance frameworks are put in place there is a real risk that situations may arise where individuals are negatively impacted because 'the computer says NO', with no recourse to meaningful explanation, correction mechanism, or way to ascertain faults that could bring about compensatory processes"

- In the case of "pervasive and automated data collection, where the individual is no longer asked to explicitly provide the data that is used by the algorithmic system, it can become difficult or impossible for individuals to identify which data were used to reach particular decision outcomes, and thus impossible to correct faulty data or assumptions, or to even ascertain if an error was made"

Vulnerability

- Vulnerability refers to person‘s heightened level of risk with regard to the experience of various harms, e.g. exploitation, exclusion, physical and social harm
- Vulnerability is a result of the interplay of the individual constellation of protective and vulnerabilty-inducing personal characteristics and the social context
- It is often linked to specific personal characteristics such as:
  - Physical characteristics that may make navigating the social world more difficult, e.g. physical disability
  - Mental characteristics that impair the ability of persons to meet normal social challenges, e.g. mental illness, cognitive disability, impaired mental capacity
  - Social characteristics that depending on a particular society may be associated with exclusion from ordinary social participation, e.g. based on race, gender, socio-economic status

Algorithmic vulnerability

Persons become subject to algorithmic decision-making

- without their knowledge and consent
- potentially with regard to highly sensitive or personally important matters (e.g. mental health risk, predictive policing, parole or credit, employment etc.)
- on the basis of potentially biased, faulty or incomplete data
- without meaningful access to redress
- Risk of self-perpetuation (one black mark keeps you in high risk category)
- once an algorithm has been developed it can be easily scaled up if data is available

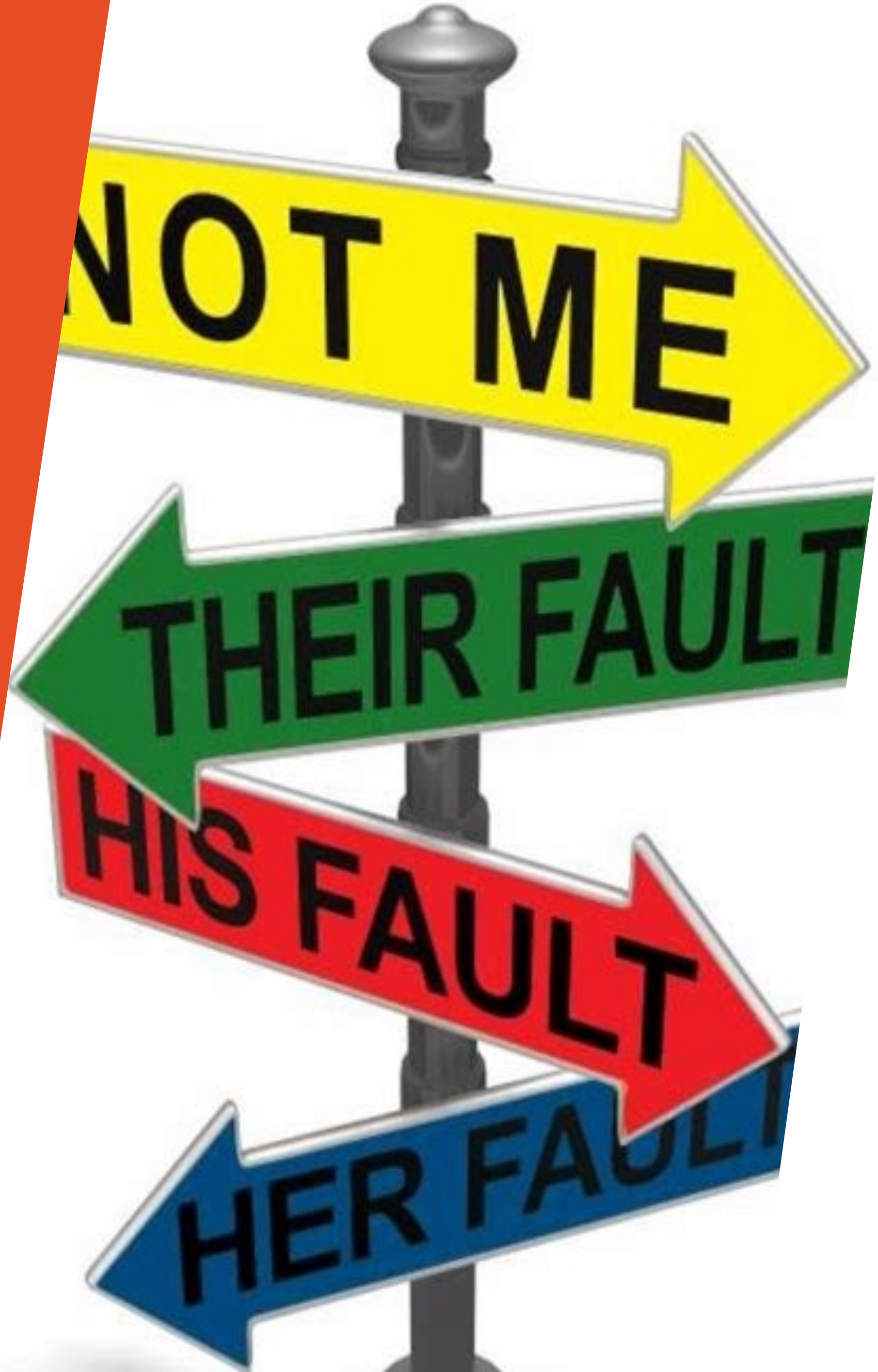
Example: automated decision-making in UK Welfare System (Guardian 2019)

- Social services meant to provide support to vulnerable persons in need of support and protection
- Automated decision-making on benefits based on drawing together wide range of data from different systems
- Systems makes decision on stopping or reducing benefits on basis of information from the system & user updates and confirmations & attendance at required meetings
- Algorithmic problems: decision at times based on faulty information, automated inference of fraudulent claims based on data, sudden impact of long forgotten historical debt, impracticable and non-negotiable automated scheduling decisions for face-to-face meetings

The problem of user-characteristics and human-in-the-loop (UK Welfare cont.)

- Structural unfairness regarding user characteristics, e.g.
  - persons with cognitive disability may lack ability to navigate complex online forms
  - Persons with mental health issues may lack ability to engage consistently with system (e.g. unable to get up, anxious, psychotic)
  - Persons who are poor or live in rural areas may not have reliable internet connection or computer
- Decreased accessibility through moving services online and implementing penalties for shortcomings in users‘ engagement with the system
- Compounded by lack of access to face-to-face support and personalised assistance
- Result: users most in need of support become disenfranchised

Algorithmic accountability



What is accountability?

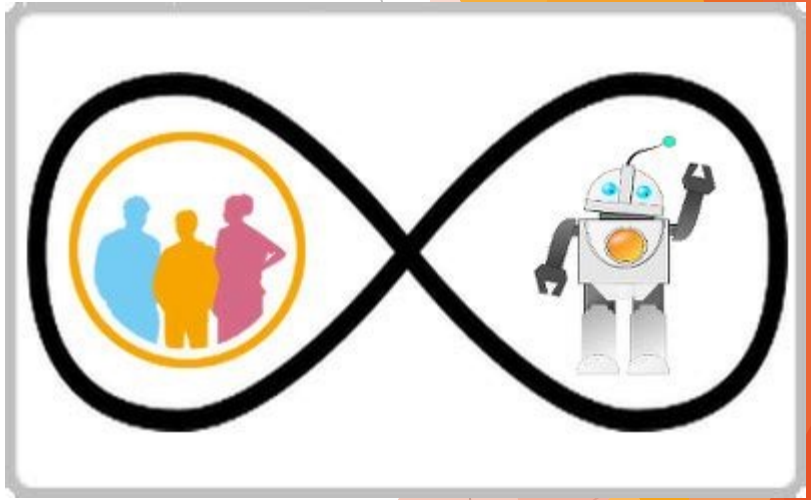
- Accountability means that persons or organisations take responsibility for their actions or failings, provide reasons for their actions, and proactively take measures to remedy any negative consequences
- Requires responsiveness to expressed concerns

Mapping challenges regarding accountability (STOA 2019)

- Complex interactions between sub-systems and data sources, not always under the control of same entity
- Unexpected outcomes that cannot be meaningfully tested for performance quality
- Difficulties in translating algorithmically derived concepts into human understandable concepts, with potential incorrect interpretations
- Information asymmetries disempowering subjects to identify incorrect results
- Accumulation of small biased decision-making, each under threshold of significant impact, but cumulatively obtaining significant impact

Accountability of algorithmic decision-making in GDPR

- GDPR includes a right NOT to be subject to automated decision-making and right to explanation
- data subject has a right to be informed, i.e. receive meaningful information about the logic involved, as well as the significance and the envisaged consequences of automated decision-making systems
- It also includes requirements to keep a "human in the loop" for any decision that have legal or significant effects for the person affected



Accountability and restorative justice

- Restorative justice focuses on addressing injustices by facilitating mutual engagement and making amends and achieving repair following injustices
- Attention to the needs and experiences of recipients with the process
- Goal is common understanding about the harm done and agreement on a way forward
- Meaningful engagement, not merely „window-dressing“
- Required at the level of organisations employing algorithmic decision-making