AI Ethics At Work

Bettina Berendt
TU Berlin, Weizenbaum Institute, & KU Leuven
VI Jobless Society Forum, Milano, 20/10/2023
Slides at www.berendt.de/bettina








Thank you, team! Pieter Delobelle, David Hartmann, Mila Miceli, Kristen Scott, Sonja Wang, Tianling Yang

Agenda

Setting the scene: AI, work, ethics, bias/fairness

AI is no monolith

Transparency is overrated

De-biasing is overrated

Interventions should strive to be systemic

- Two examples from our academic action space
- Larger alliances

AI and work – concerns and hopes

https://www.fortuneita.com/2023/08/06/mckinsey-intelligenza-artificiale-stravolgera-la-forza-lavoro/
https://futuranetwork.eu/ricerca-e-tecnologia/707-3815/intelligenza-artificiale-e-discriminazione-di-genere-quali-sfide-dovremo-affrontare
https://invaliditaediritti.it/lavoro-per-disabili-con-intelligenza-artificiale/
https://www.noemamag.com/the-exploited-labor-behind-artificial-intelligence/

EXISTENCE of jobs

ACCESS to jobs and related resources

Working CONDITIONS

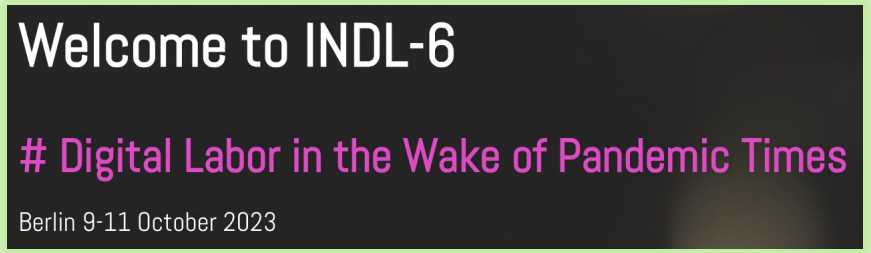
McKinsey, l'intelligenza artificiale stravolgerà la forza lavoro

Intelligenza artificiale e discriminazione di genere

Lavoro per disabili con intelligenza artificiale

The Exploited Labor Behind Artificial Intelligence

Supporting transnational worker organizing should be at the center of the fight for "ethical AI."



Ethics and “bias” (aka unfairness)

Ethics: How to act

What to build ?
How to explain ?
How to deal with limitations ?

- in a good way – e.g. so as to maximise happiness/pleasure/utility/…
- in a right way – e.g. so as to respect human rights (e.g. privacy)
- in a fair and unbiased way that does not discriminate
- …

Technology unfairness: example problems

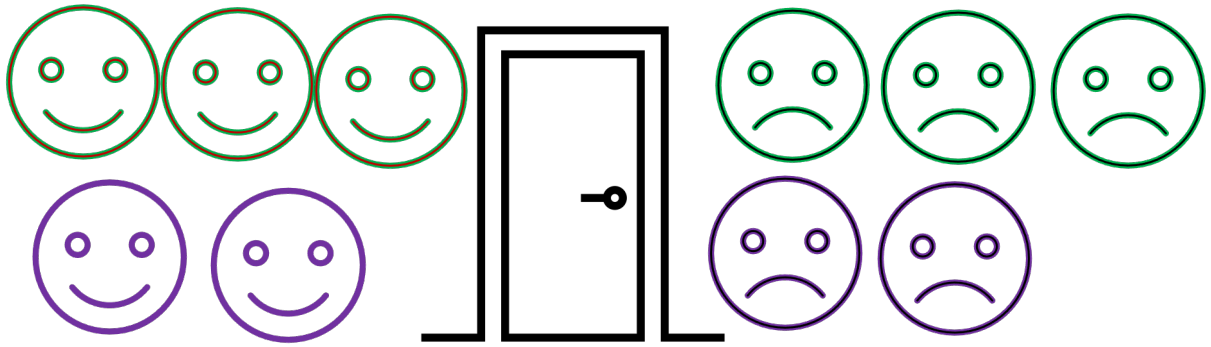


Automated decisions that discriminate

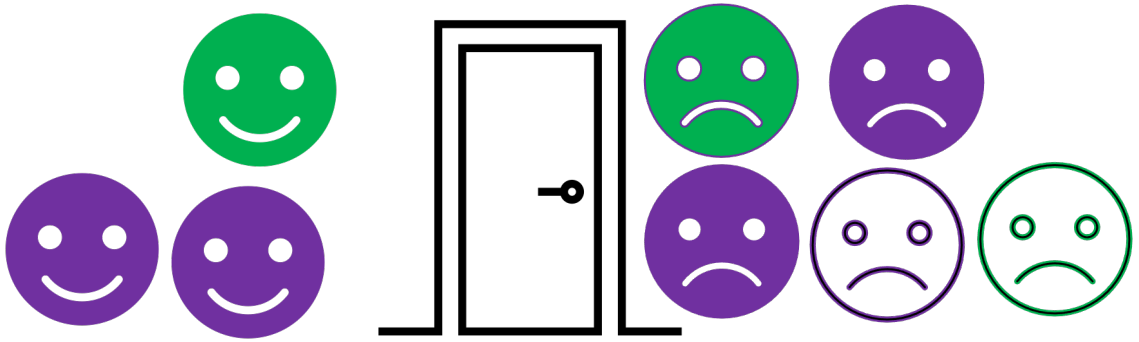
- in recruiting and hiring
- in support given to jobseekers
- in recommendations given to jobseekers
- …

Technology fairness goals: 2 example definitions

"demographic parity"



"equal opportunity"



For more definitions, see Castelnovo, A., Crupi, R., Greco, G. et al. A clarification of the nuances in the fairness metrics landscape. Sci Rep 12, 4209 (2022). https://doi.org/10.1038/s41598-022-07939-1

AI is no monolith

Narayanan: How to recognize AI Snake Oil (2019)

https://www.cs.princeton.edu/~arvindn/talks/MIT-STS-AI-snakeoil.pdf

Genuine, rapid progress

- Shazam, reverse img search
- Face recognition
- Med. diagnosis from scans
- Speech to text
- Deepfakes

Perception

Ethical concerns

because of high accuracy

Narayanan: How to recognize AI Snake Oil (2019)

https://www.cs.princeton.edu/~arvindn/talks/MIT-STS-AI-snakeoil.pdf

Genuine, rapid progress	**Imperfect but improving**
• Shazam, reverse img search	• Spam detection
• Face recognition	• Copyright violation
• Med. diagnosis from scans	• Automated essay grading
• Speech to text	• Hate speech detection
• Deepfakes	• Content recommendation

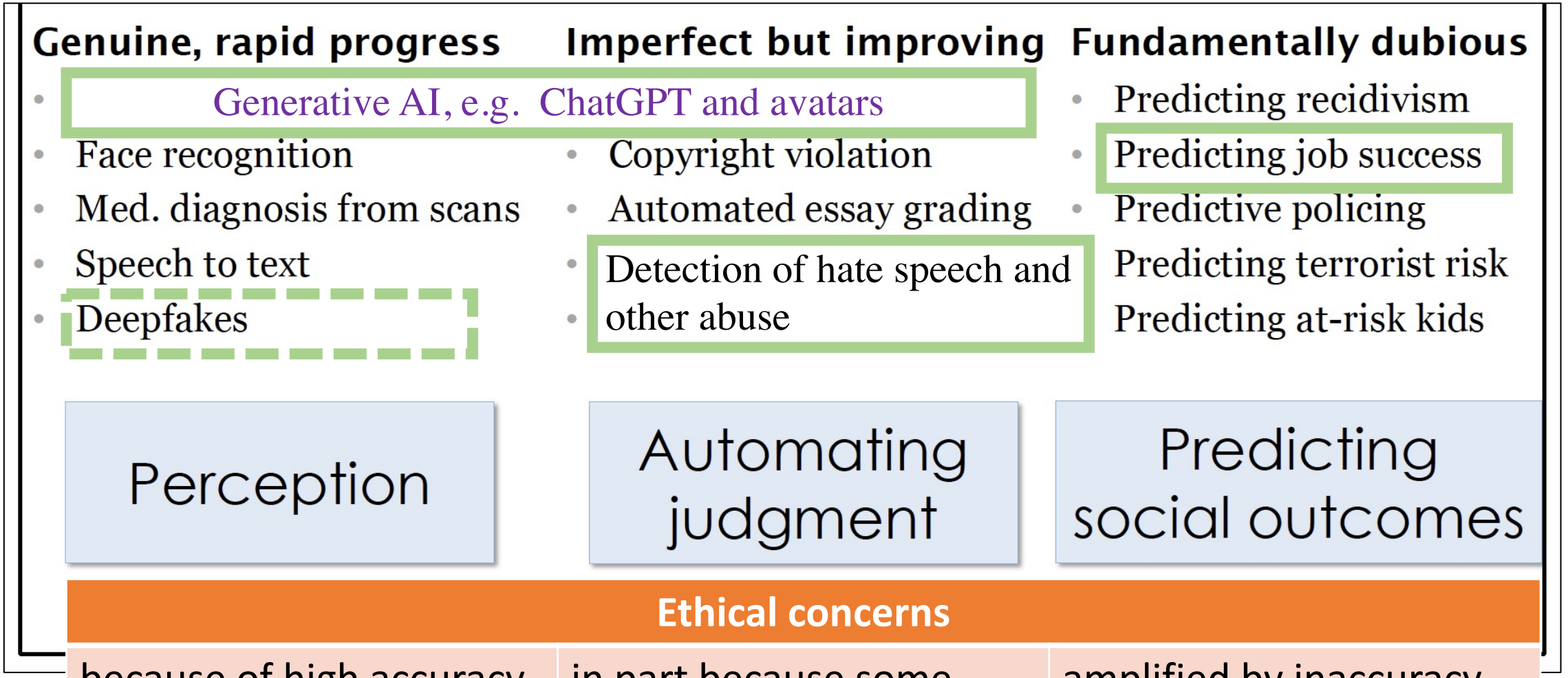
Perception	Automating judgment
Ethical concerns	
because of high accuracy	in part because some error is inevitable

Narayanan: How to recognize AI Snake Oil (2019)

https://www.cs.princeton.edu/~arvindn/talks/MIT-STS-AI-snakeoil.pdf

Genuine, rapid progress	Imperfect but improving	Fundamentally dubious
• Shazam, reverse img search	• Spam detection	• Predicting recidivism
• Face recognition	• Copyright violation	• Predicting job success
• Med. diagnosis from scans	• Automated essay grading	• Predictive policing
• Speech to text	• Hate speech detection	• Predicting terrorist risk
• Deepfakes	• Content recommendation	• Predicting at-risk kids
Perception	Automating judgment	Predicting social outcomes
Ethical concerns		
because of high accuracy	in part because some error is inevitable	amplified by inaccuracy

Work-related applications of AI (examples)

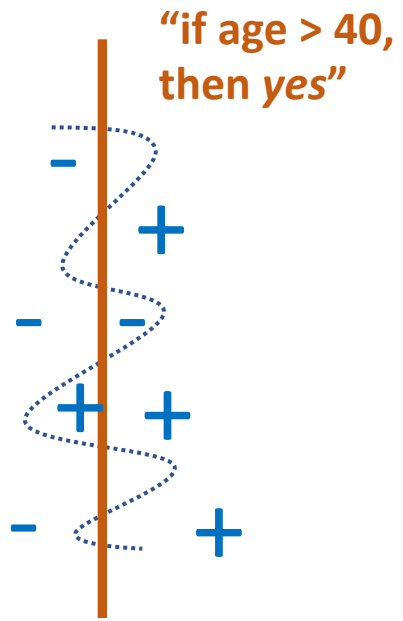


Transparency is overrated

Transparency = “opening the black box”

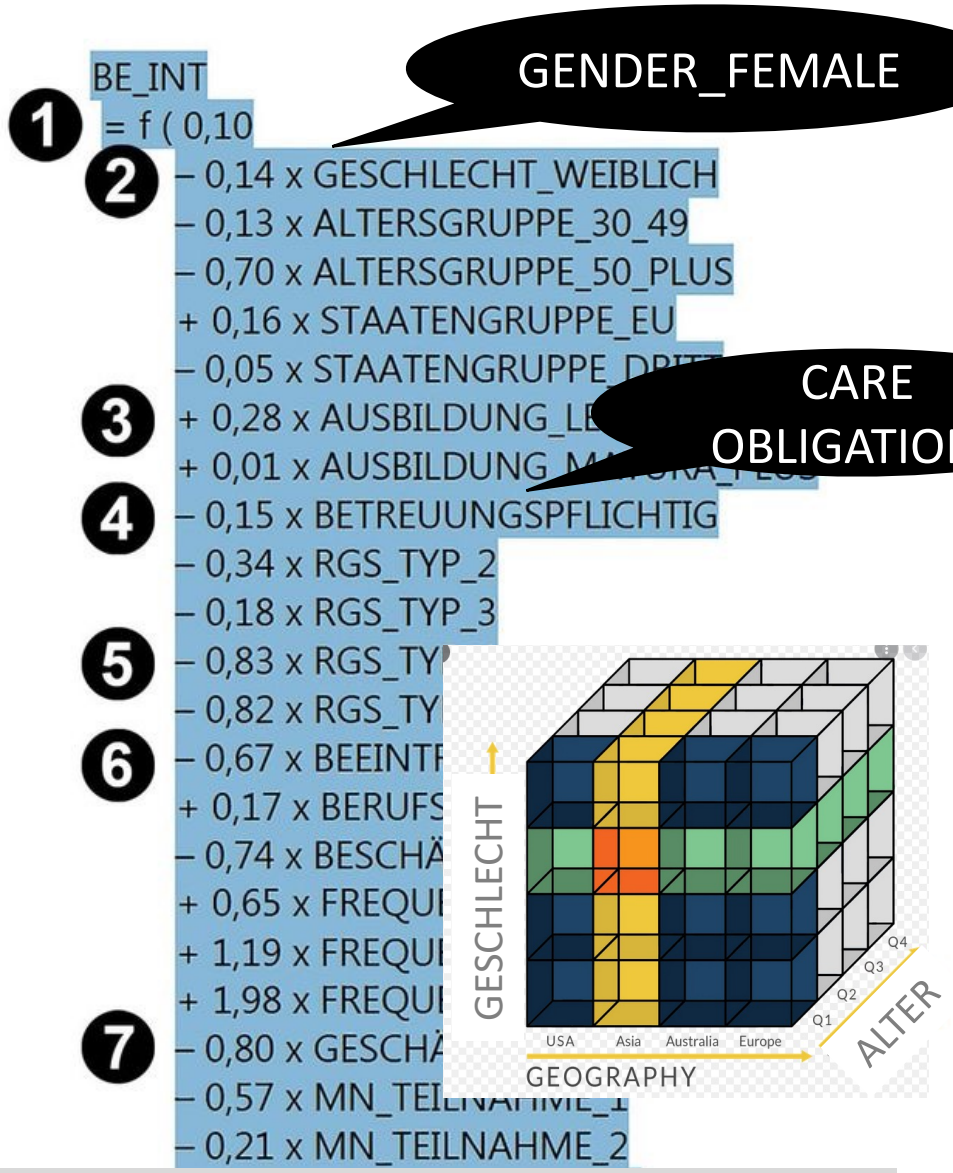
Lots of computer-science research

- Opacity a frequent complaint
- Transparency / explanations / etc. a frequently invoked remedy
- A cornerstone of EU AI Act proposal



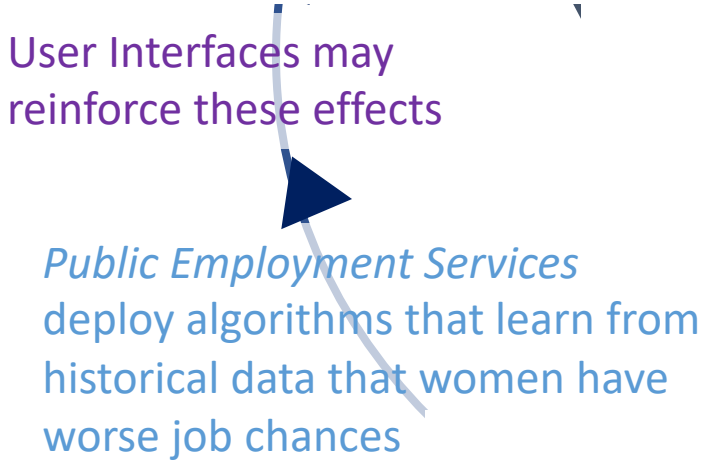
Transparency *can* help

- Ex.: “Job market chances system” of the Austrian Public Employment Service (AMS)
- Purpose: predict future chances of jobseekers based on statistics from past years.
- Jobseekers are classified into three groups according to their “integration chance”.
  - Descriptive: Results “just reflect the harsh realities of the labour market“.
  - Prescriptive: The “middle“ group is the focus for allocating support (esp. training measures).
- *Highly* controversial
- Status since 2021: AMS vs. data protection authority – Federal Administrative Court, no judgment yet



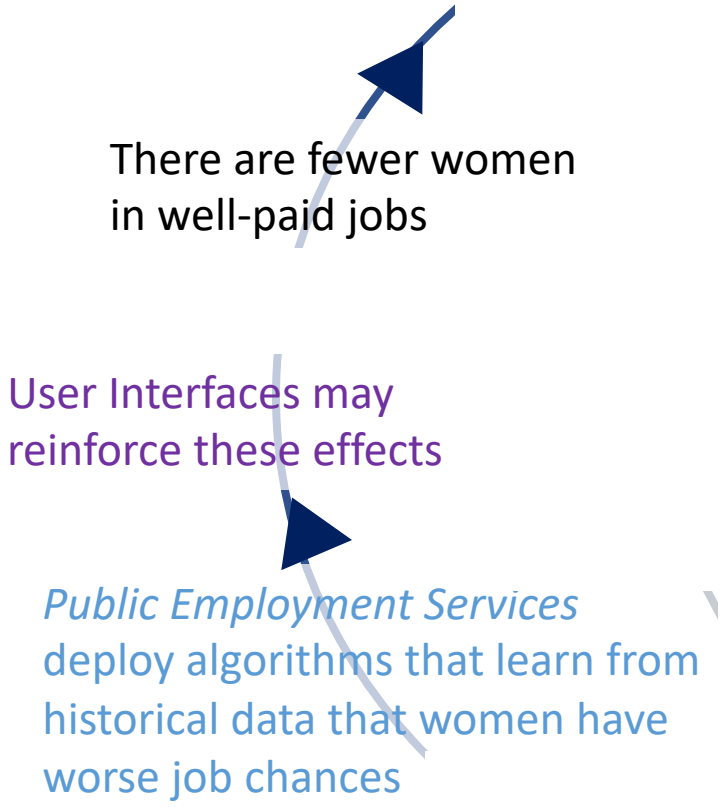
Doris Allhutter, Astrid Mager, Florian Cech, Fabian Fischer, and Gabriel Grill. 2020. Der AMS Algorithmus - Eine Soziotechnische Analyse des Arbeitsmarktchancen-Assistenz-Systems (AMAS). Technical Report. Wien: ÖAW, ITA 2020-02. http://epub.oeaw.ac.at/?arp=0x003bfdf3/

But there’s not only 1 decision! Pernicious feedback loops



Berendt, B. (2022). The AI Act Proposal: Towards the next transparency fallacy? Why AI regulation should be based on principles based on how algorithmic discrimination works. In BMJ & F. Rostalski (Eds.), *Künstliche Intelligenz - Wie gelingt eine vertrauenswürdige Verwendung in Deutschland und Europa?* (pp. 31-52). Tübingen, Germany: Mohr Siebeck. https://people.cs.kuleuven.be/~bettina.berendt/Papers/berendt_2022c_last_author_version.pdf

But there’s not only 1 decision! Pernicious feedback loops



Berendt, B. (2022). The AI Act Proposal: Towards the next transparency fallacy? Why AI regulation should be based on principles based on how algorithmic discrimination works. In BMJ & F. Rostalski (Eds.), *Künstliche Intelligenz - Wie gelingt eine vertrauenswürdige Verwendung in Deutschland und Europa?* (pp. 31-52). Tübingen, Germany: Mohr Siebeck. https://people.cs.kuleuven.be/~bettina.berendt/Papers/berendt_2022c_last_author_version.pdf

But there's not only 1 decision! Pernicious feedback loops

People perceive a world in which it is men who tend to have the well-paid jobs (observation a/o via media/texts)

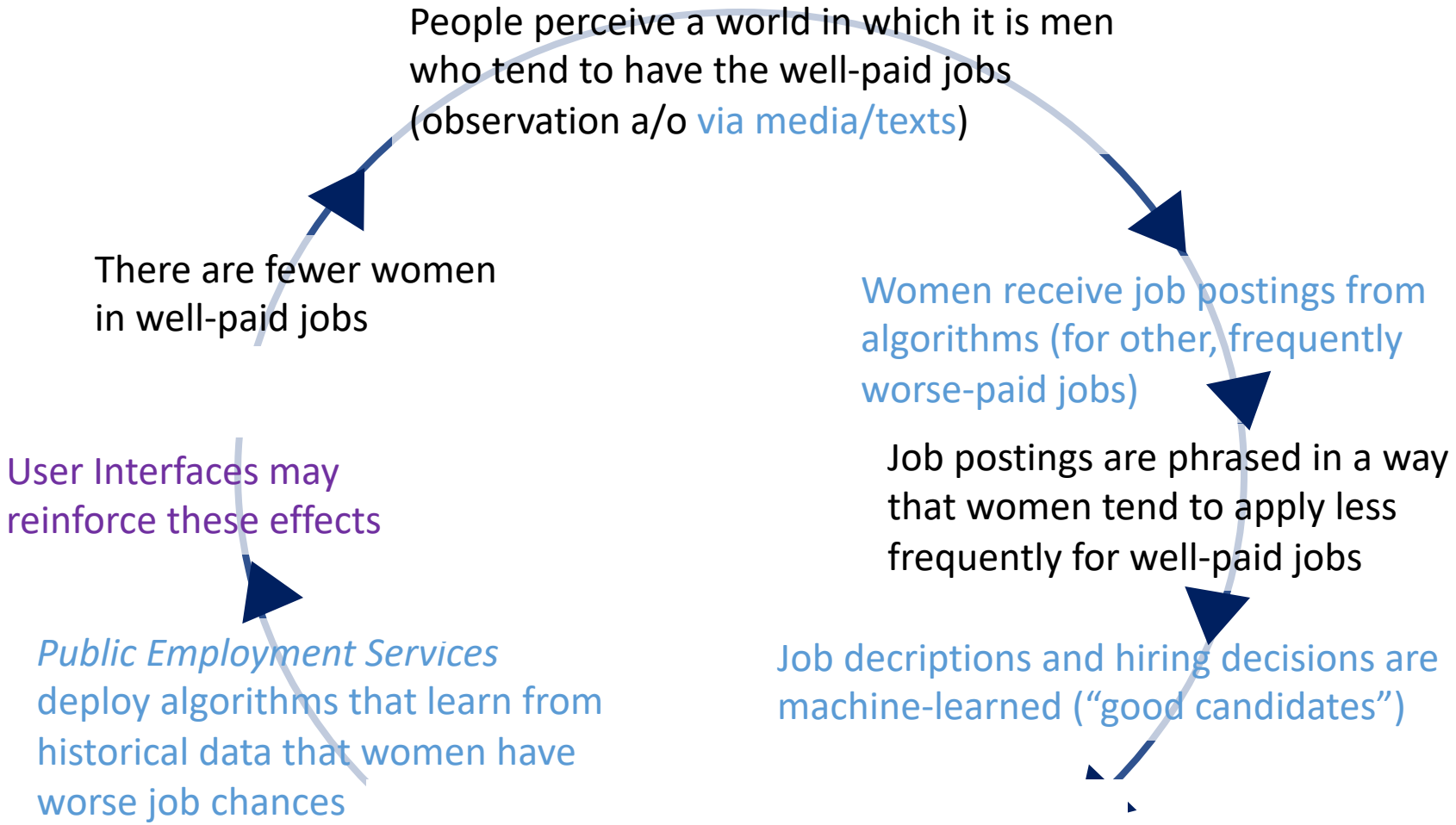
There are fewer women in well-paid jobs

User Interfaces may reinforce these effects

Public Employment Services deploy algorithms that learn from historical data that women have worse job chances

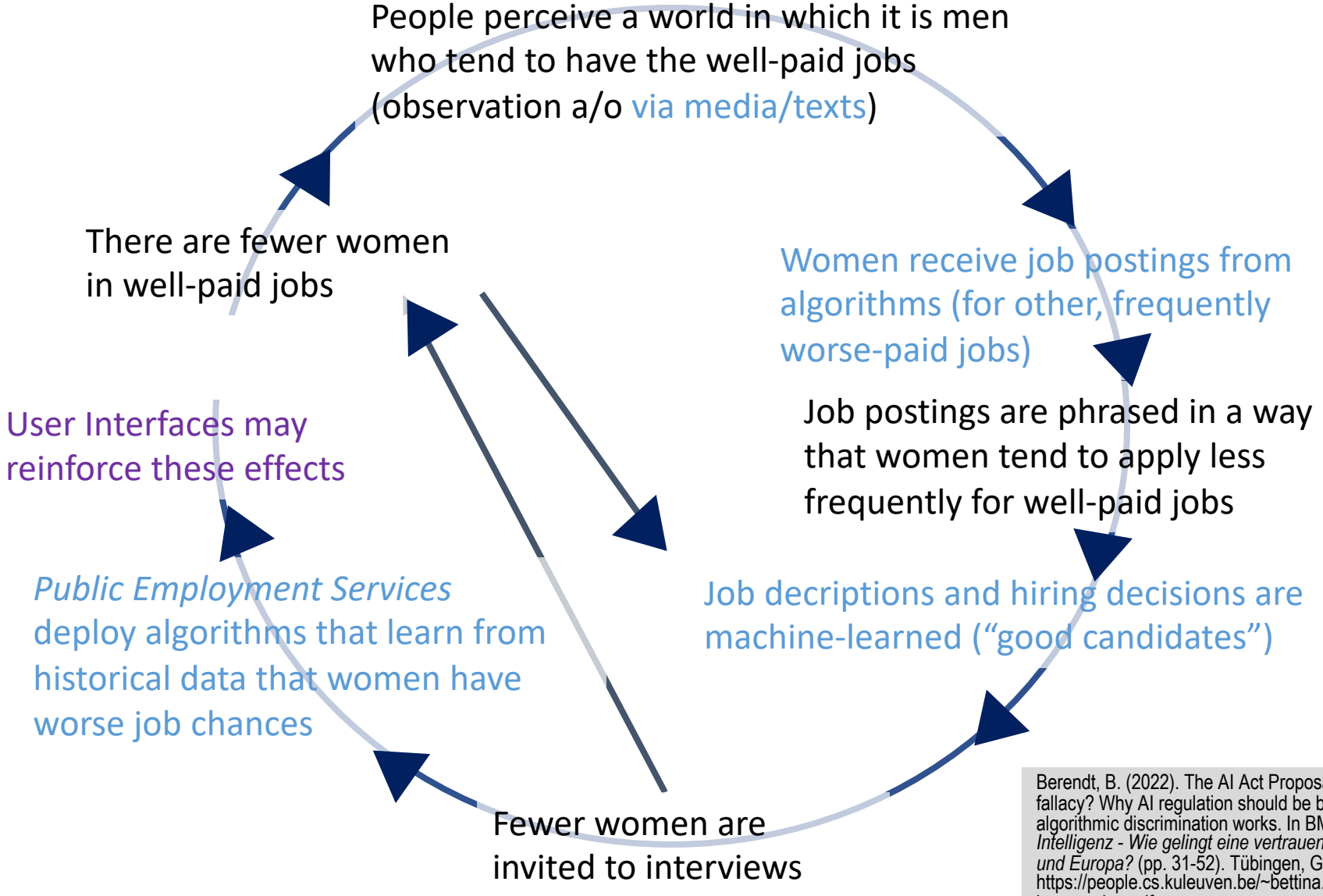
Berendt, B. (2022). The AI Act Proposal: Towards the next transparency fallacy? Why AI regulation should be based on principles based on how algorithmic discrimination works. In BMJ & F. Rostalski (Eds.), *Künstliche Intelligenz - Wie gelingt eine vertrauenswürdige Verwendung in Deutschland und Europa?* (pp. 31-52). Tübingen, Germany: Mohr Siebeck. https://people.cs.kuleuven.be/~bettina.berendt/Papers/berendt_2022c_last_author_version.pdf

But there's not only 1 decision! Pernicious feedback loops



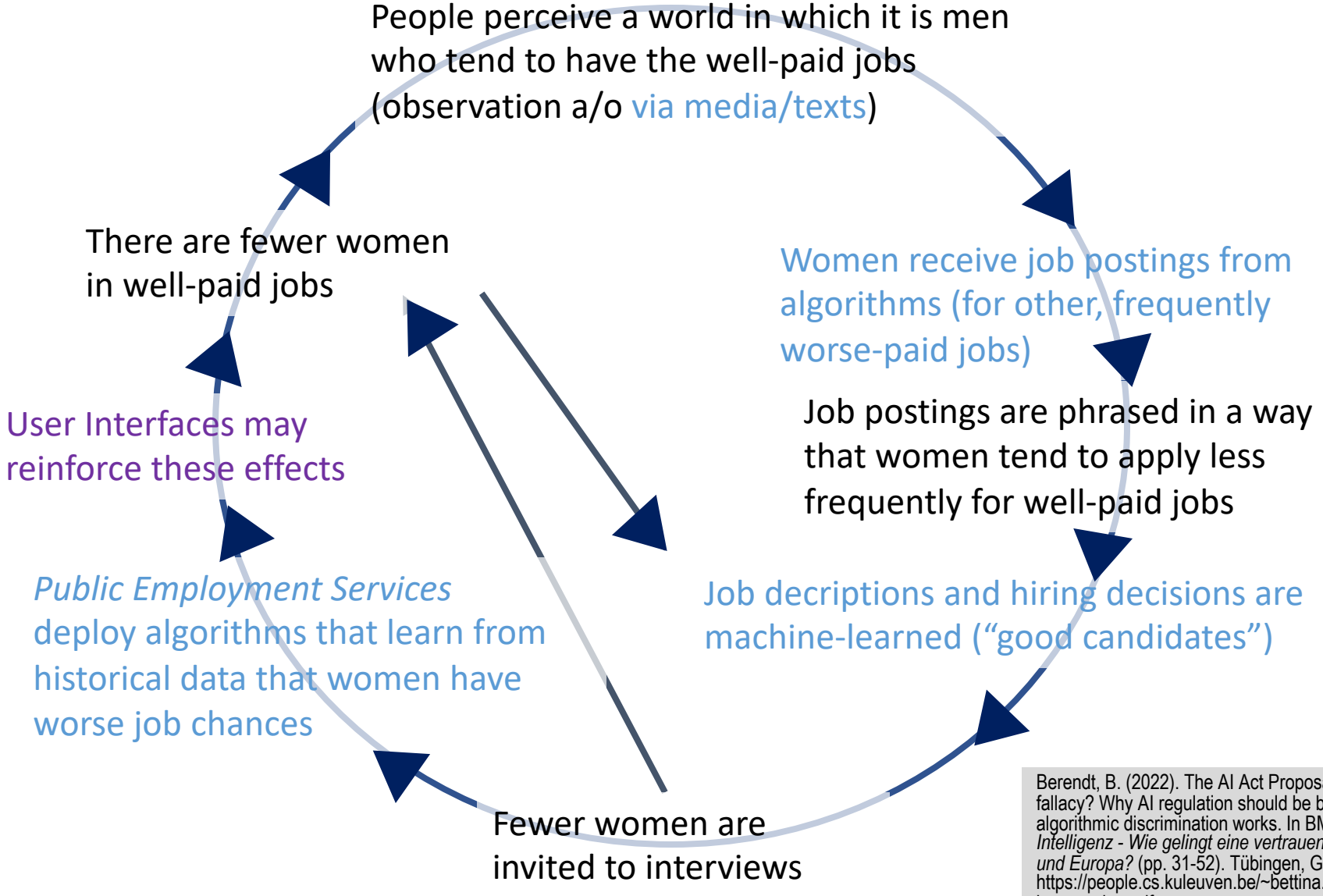
Berendt, B. (2022). The AI Act Proposal: Towards the next transparency fallacy? Why AI regulation should be based on principles based on how algorithmic discrimination works. In BMJ & F. Rostalski (Eds.), *Künstliche Intelligenz - Wie gelingt eine vertrauenswürdige Verwendung in Deutschland und Europa?* (pp. 31-52). Tübingen, Germany: Mohr Siebeck. https://people.cs.kuleuven.be/~bettina.berendt/Papers/berendt_2022c_last_author_version.pdf

But there’s not only 1 decision! Pernicious feedback loops



Berendt, B. (2022). The AI Act Proposal: Towards the next transparency fallacy? Why AI regulation should be based on principles based on how algorithmic discrimination works. In BMJ & F. Rostalski (Eds.), *Künstliche Intelligenz - Wie gelingt eine vertrauenswürdige Verwendung in Deutschland und Europa?* (pp. 31-52). Tübingen, Germany: Mohr Siebeck. https://people.cs.kuleuven.be/~bettina.berendt/Papers/berendt_2022c_last_author_version.pdf

But there's not only 1 decision! Pernicious feedback loops



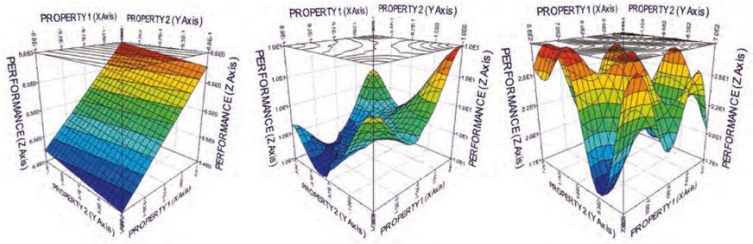
Berendt, B. (2022). The AI Act Proposal: Towards the next transparency fallacy? Why AI regulation should be based on principles based on how algorithmic discrimination works. In BMJ & F. Rostalski (Eds.), *Künstliche Intelligenz - Wie gelingt eine vertrauenswürdige Verwendung in Deutschland und Europa?* (pp. 31-52). Tübingen, Germany: Mohr Siebeck. https://people.cs.kuleuven.be/~bettina.berendt/Papers/berendt_2022c_last_author_version.pdf

De-biasing is overrated

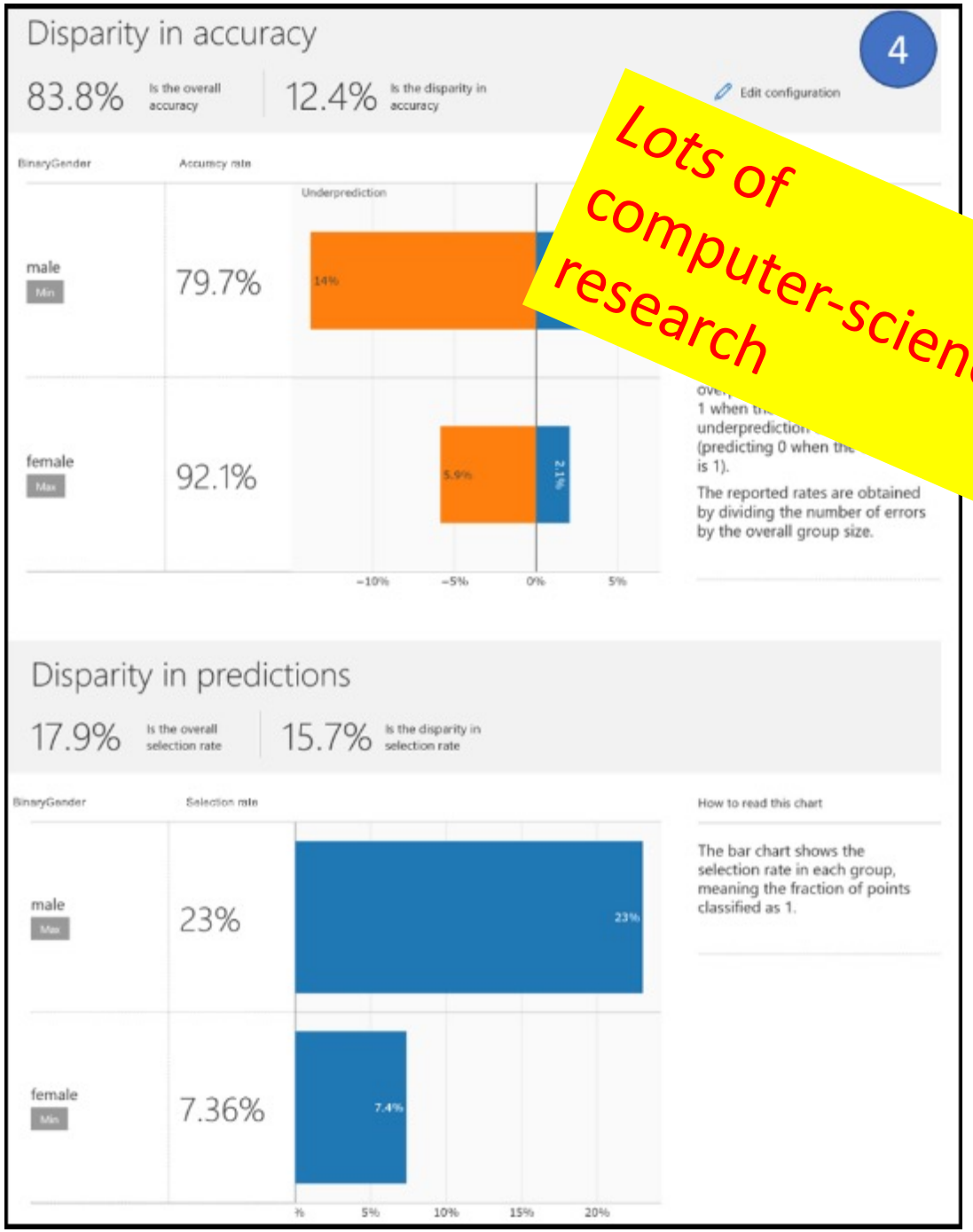
De-biasing algorithms

Ex.: Fairlearn package, used in Microsoft Azure Machine Learning

A form of multi-objective optimisation

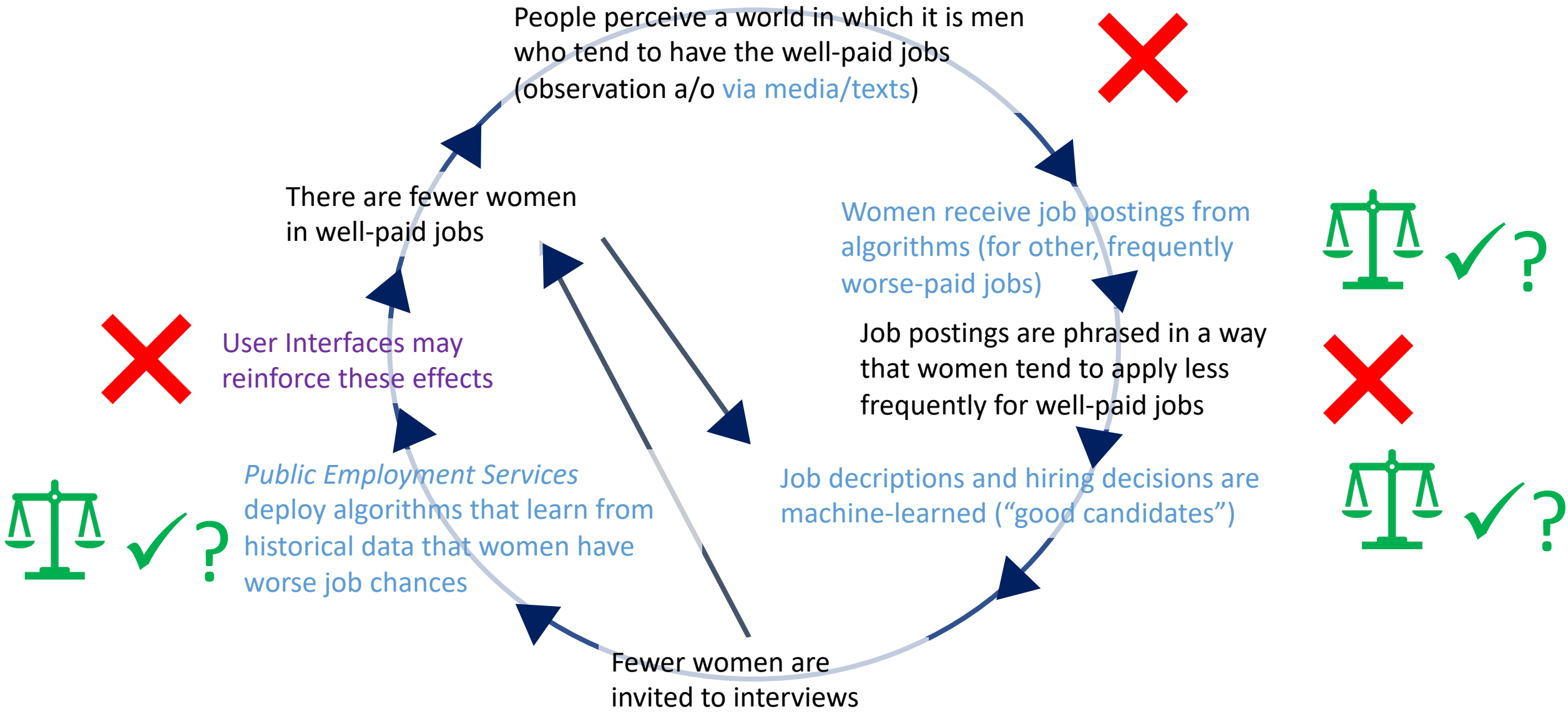


https://opendatascience.com/how-to-assess-ai-systems-fairness-and-mitigate-any-observed-unfairness-issues/
https://parametrichouse.com/optimizing-creatively-in-multi-objective-optimization/



Lots of computer-science research

De-biasing *can* help – but only so far



Increasing amounts of research

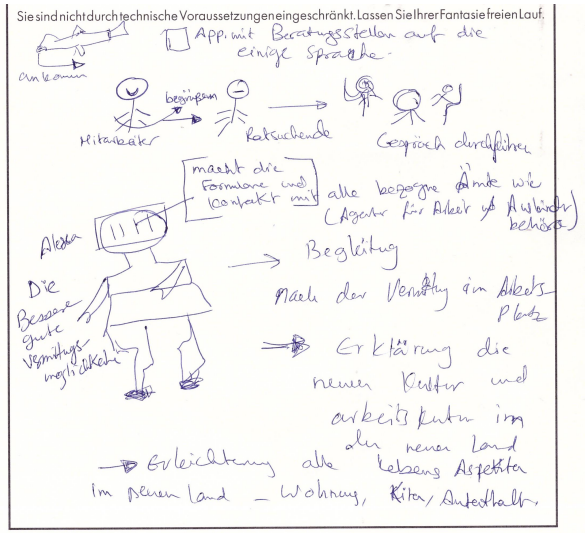
More systemic interventions I: from academia

(2 examples from our own work)

Intervention ex. 1: Co-design & co-creation for jobseekers

From co-design workshops with stakeholders (jobseekers, counsellors)





Kristen M. Scott, Sonja Mei Wang, Milagros Miceli, Pieter Delobelle, Karolina Sztandar-Sztanderska, Bettina Berendt: Algorithmic Tools in Public Employment Services: Towards a Jobseeker-Centric Perspective. FAccT 2022: 2138-2148. https://dl.acm.org/doi/10.1145/3531146.3534631

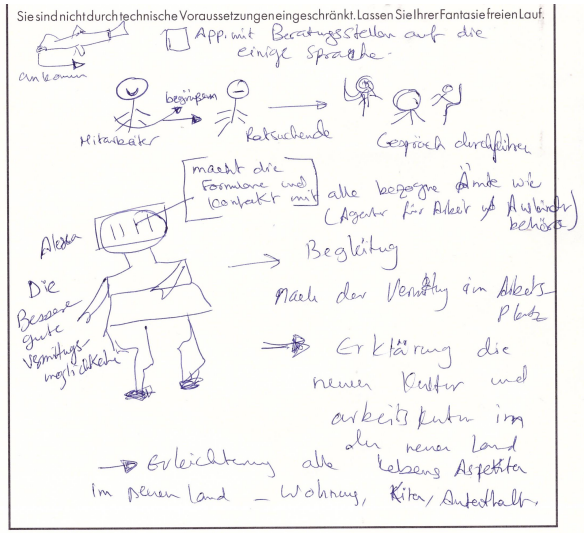
Sonja Mei Wang, Kristen M Scott, Margarita Artemenko, Milagros Miceli and Bettina Berendt.``We try to empower them'' - Exploring future technologies to support migrant jobseekers. FAccT 2023: 972-983. https://dl.acm.org/doi/10.1145/3593013.3594056

Pieter Delobelle, Sonja Mei Wang, Bettina Berendt (2023). ResumeTailor: Improving Resume Quality Through Co-Creative Tools. Frontiers in Artificial Intelligence and Applications, Volume 368: HHAI 2023: Augmenting Human Intellect, 419-421. https://ebooks.iospress.nl/volumearticle/63364

Intervention ex. 1: Co-design & co-creation for jobseekers

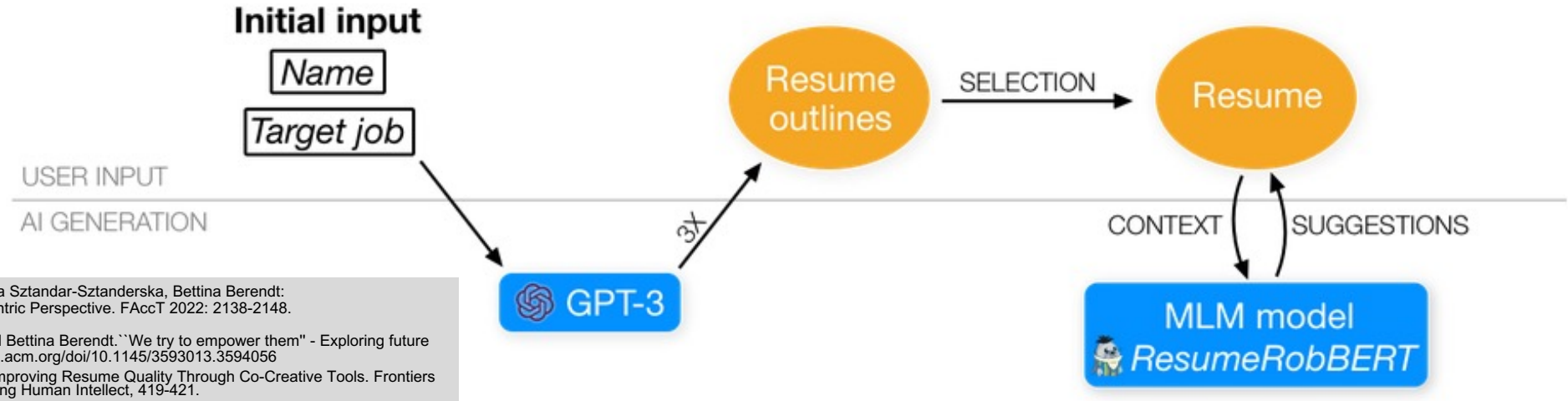
From co-design workshops with stakeholders (jobseekers, counsellors) to co-creative tool for resumé writing







ResumeTailor



Kristen M. Scott, Sonja Mei Wang, Milagros Miceli, Pieter Delobelle, Karolina Sztandar-Sztanderska, Bettina Berendt: Algorithmic Tools in Public Employment Services: Towards a Jobseeker-Centric Perspective. FAccT 2022: 2138-2148. https://dl.acm.org/doi/10.1145/3531146.3534631

Sonja Mei Wang, Kristen M Scott, Margarita Artemenko, Milagros Miceli and Bettina Berendt.``We try to empower them'' - Exploring future technologies to support migrant jobseekers. FAccT 2023: 972-983. https://dl.acm.org/doi/10.1145/3593013.3594056

Pieter Delobelle, Sonja Mei Wang, Bettina Berendt (2023). ResumeTailor: Improving Resume Quality Through Co-Creative Tools. Frontiers in Artificial Intelligence and Applications, Volume 368: HHAI 2023: Augmenting Human Intellect, 419-421. https://ebooks.iospress.nl/volumearticle/63364

Intervention ex. 2:
Teaching AI ethics with ChatGPT

- Preparing students for their (later) work as computer-science professionals
- Use *and* reflect *and* be critical (in a methodological way) about ChatGPT

www.berendt.de/teaching.html

Increasing amounts of practice

More systemic interventions II:
Larger alliances

AI Snake Oil & workers' rights: a further twist



TIME

SPOTLIGHT STORY MAYA MOORE AND JONATHAN IRONS ON THEIR LOVE STORY

BUSINESS • TECHNOLOGY

Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic

BY **BILLY PERRIGO** JANUARY 18, 2023 7:00 AM EST

https://tutanota.com/blog/chat-control-criticism
https://time.com/6247678/openai-chatgpt-kenya-workers/
https://edri.org/our-work/most-criticised-eu-law-of-all-time/

AI Snake Oil & workers' rights: a further twist

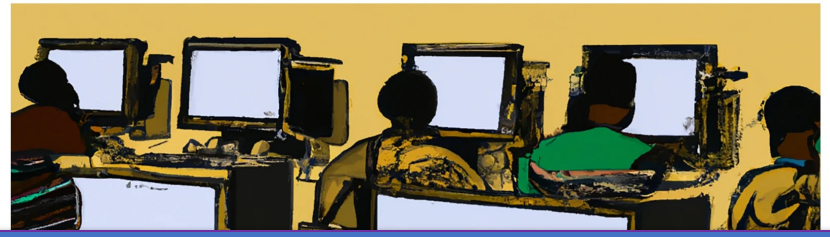
TIME

SPOTLIGHT STORY MAYA MOORE AND JONATHAN IRONS ON THEIR LOVE STORY

BUSINESS • TECHNOLOGY

Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic

BY **BILLY PERRIGO** JANUARY 18, 2023 7:00 AM EST



Chat Control Criticism: Why the EU CSAM Scanning Plans Must Fail.

The EU Regulation to Prevent and Combat Child Sexual Abuse has become the "most criticized law of all time".

Idea: client-side scanning, AI classification of CSAM -> surveillance; automating judgment -> dubious!

https://tutanota.com/blog/chat-control-criticism
https://time.com/6247678/openai-chatgpt-kenya-workers/
https://edri.org/our-work/most-criticised-eu-law-of-all-time/

AI Snake Oil & workers' rights: a further twist



TIME

SPOTLIGHT STORY MAYA MOORE AND JONATHAN IRONS ON THEIR LOVE STORY

BUSINESS • TECHNOLOGY

Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic

BY BILLY PERRIGO JANUARY 18, 2023 7:00 AM EST

"We do not need to worry too much about protecting data workers' rights, since their jobs will be taken over by AI anyway."

Chat Control Criticism: Why the EU CSAM Scanning Plans Must Fail.

The EU Regulation to Prevent and Combat Child Sexual Abuse has become the "most criticized law of all time".

Idea: client-side scanning, AI classification of CSAM -> surveillance; automating judgment -> dubious!

https://tutanota.com/blog/chat-control-criticism
https://time.com/6247678/openai-chatgpt-kenya-workers/
https://edri.org/our-work/most-criticised-eu-law-of-all-time/

AI Snake Oil & workers' rights: a further twist


TIME

SPOTLIGHT STORY MAYA MOORE AND JONATHAN IRONS ON THEIR LOVE STORY

BUSINESS • TECHNOLOGY

Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic

BY BILLY PERRIGO JANUARY 18, 2023 7:00 AM EST

Chat Control Criticism: Why the EU CSAM Scanning Plans Must Fail.

The EU Regulation to Prevent and Combat Child Sexual Abuse has become the "most criticized law of all time".

Idea: client-side scanning, AI classification of CSAM -> surveillance; automating judgment -> dubious!

"We do not need to worry too much about protecting data workers' rights, since their jobs will be taken over by AI anyway."

Interventions:
Scientific analyses and alliances with legal & political.

Final EU Council vote postponed till late Oct. 2023

Abelson, H., Anderson, R., Bellovin, S.M., Benaloh, J., Blaze, M., Callas, J., Diffie, W., Landau, S., Neumann, P.G., Rivest, R.L., Schiller, J.I., Schneier, B., Teague, V., & Troncoso, C. (2021). *Bugs in our pockets: The risks of client-side scanning.* https://arxiv.org/abs/2110.07450

https://tutanota.com/blog/chat-control-criticism
https://time.com/6247678/openai-chatgpt-kenya-workers/
https://edri.org/our-work/most-criticised-eu-law-of-all-time/

https://www.reuters.com/legal/litigation/us-copyright-office-denies-protection-another-ai-created-image-2023-09-06/
https://www.theguardian.com/commentisfree/2023/oct/07/screenwriters-won-a-historic-victory-against-ai-the-rest-of-us-should-follow
https://projection-mapping.org/narendra-modis-holograms/
https://www.spiegel.de/kultur/musik/abba-voyage-in-london-abba-als-avatare-agnethas-augen-glaenzen-als-koennten-sie-echte-traenen-weinen-a-1075b60e-756d-4a5a-9922-751e3665196a
https://www.youtube.com/watch?v=jjJv-BF_UQE

Economic aspects

US Copyright Office denies protection for another AI-created image 6 Sep 2023

7 Oct 2023: WGA Union wins contract guarantees that AI is not considered a “writer”, that companies cannot force writers to use AI, and that companies must disclose if writers are given any AI-generated material to work with.

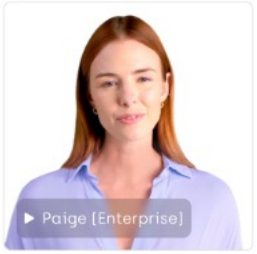
Ongoing: Fight of Sag-Aftra union: background actors as one-off raw material for re-usable avatars

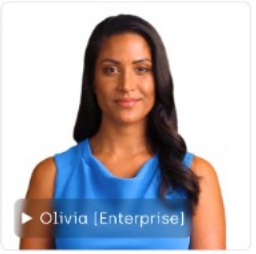
Same technology, different economic power over it ➔ harm or benefit ?











Legal aspects: Balancing different rights

- Client-side scanning
  - privacy/data protection ←→ safety
  - protection from psychological harm and exploitation ←→ ~
- Authors & actors
  - autonomy ←→ copyright, personality rights, economic vulnerability
- Many other balancing acts, e.g. detailed automated image recognition:
  - privacy/data protection ←→ information accessibility for visually impaired people *

* Higgins (2023). The many faces of equality. Talk at *Unesco AI4IA Conference*. https://ai4iaconference.com/

In summary: Take-home messages

Setting the scene: AI, work, ethics, bias/fairness

AI is no monolith

Transparency is overrated

De-biasing is overrated

Interventions should strive to be systemic

- Two examples from our academic action space
- Larger alliances

Thank you!
Looking forward to the discussion
and learning about our blind spots!