

Privacy, fairness, diversity and accountability: limits of formalising social goals

Bettina Berendt
https://people.cs.kuleuven.be/~bettina.berendt

KU LEUVEN DTAI

SocInfo 2018, 28 Sep 2018, St. Petersburg

Terms are not neutral. Neither are concepts.

- Expat?
- Economic migrant?

Computers, Big Data, AI and social goals/values

Who's being discriminated against here?



Ad related to **latanya sweeney** ⓘ

Latanya Sweeney Truth
www.instantcheckmate.com/
Looking for **Latanya Sweeney**? Check **Latanya Sweeney's** Arrests.

Ads by Google

Latanya Sweeney, Arrested?
1) Enter Name and State. 2) Access Full Background Checks Instantly.
www.instantcheckmate.com/

Latanya Sweeney
Public Records Found For: **Latanya Sweeney**. View Now.
www.publicrecords.com/

La Tanya
Search for La Tanya Look Up Fast Results now!
www.ask.com/La+Tanya

(Sweeney, 2013)

Formal measures: examples

Formalising fairness: Discrimination-aware / fairness-aware data mining

PD and PND items: potentially (not) discriminatory

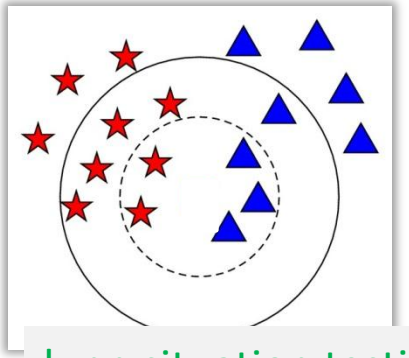
goal: detect & block mined rules such as

purpose=new_car & gender = female → credit=no

measures of discriminatory power of a rule include

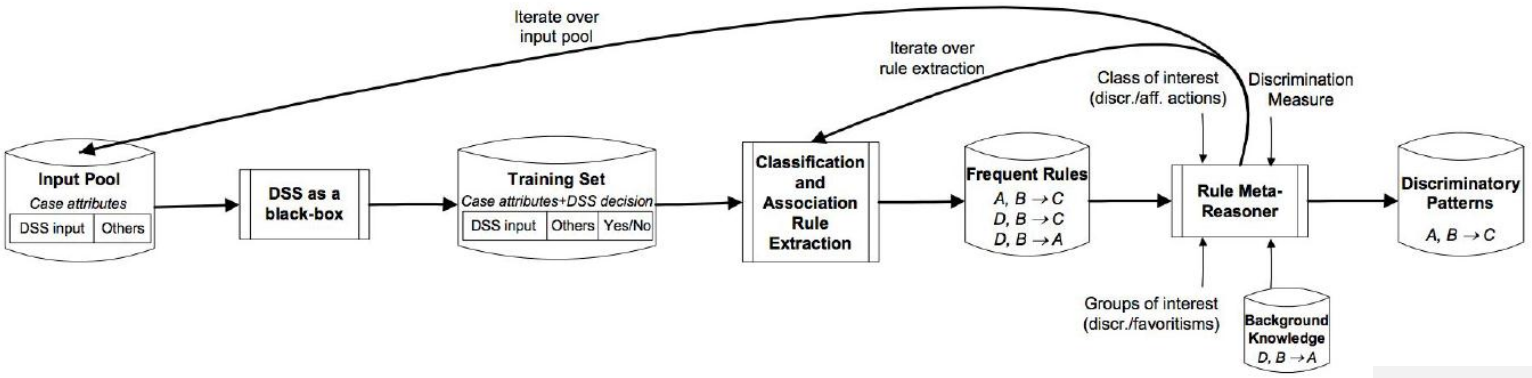
elift (B&A → C) = conf (B&A → C) / conf (B → C) ,

where A is a PD item and B a PND item



k-nn situation testing

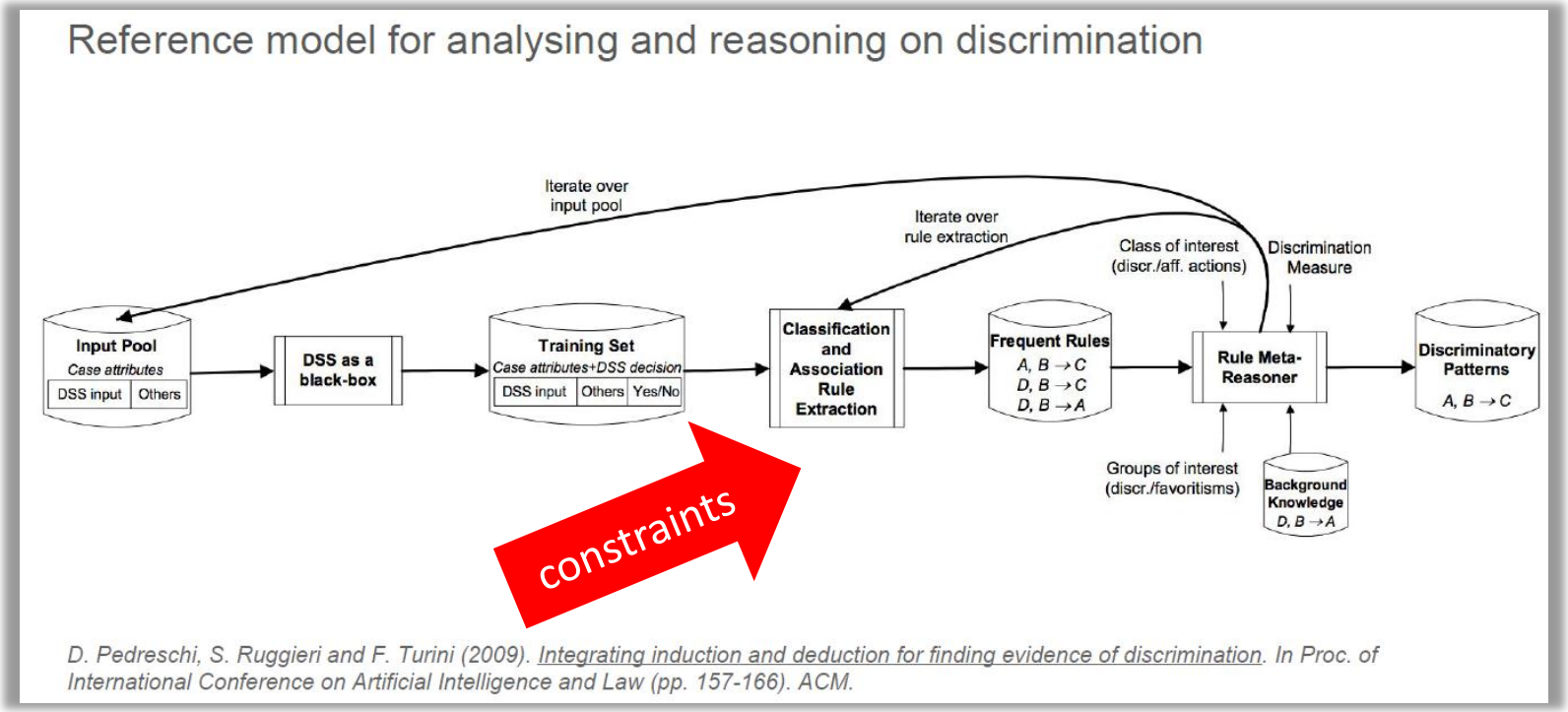
Reference model for analysing and reasoning on discrimination



D. Pedreschi, S. Ruggieri and F. Turini (2009). Integrating induction and deduction for finding evidence of discrimination. In Proc. of International Conference on Artificial Intelligence and Law (pp. 157-166). ACM.

(Pedreschi et al., 2008, 2009; Thanh et al., 2011)

Discrimination-aware / fairness-aware data mining: From detection to prevention



Formalising privacy

K-Anonymity

A 3-anonymous patient table

Zipcode	Age	Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
4790*	≥40	Flu
4790*	≥40	Heart Disease
4790*	≥40	Cancer
476**	3*	Heart Disease
476**	3*	Cancer
476**	3*	Cancer

Formalising privacy

Attacks Against K-Anonymity

- **k-Anonymity does not provide privacy if:**
  - **Sensitive values in an equivalence class lack diversity**
  - **The attacker has background knowledge**

Homogeneity Attack

Bob	
Zipcode	***Age***
47678	27

Background Knowledge Attack

Umeko (Japanese)	
Zipcode	***Age***
47673	36

A 3-anonymous patient table

Zipcode	Age	Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
4790*	≥40	Flu
4790*	≥40	Heart Disease
4790*	≥40	Cancer
476**	3*	Heart Disease
476**	3*	Cancer
476**	3*	Cancer

(Slide by Murat Kantarcioglu)

Formalising diversity

Source Diversity	Content Diversity	Exposure Diversity
1. Ownership	1. Program-Type Format	1. Horizontal
a. Programming	2. Demographic	2. Vertical
b. Outlet	3. Idea/Viewpoint	
2. Workforce		

Figure 1. Diversity components, subcomponents, and assumed relationships

Table 2. Four variants of Δ and links with diversity properties and measures.

property	α	β	equation (9.2): $\Delta =$	equivalents (cf. table 1)	interpretation
variety	0	0	$\sum_{ij} d_{ij}^0$	$(N^2 - N)/2$	scaled variety
balance	0	1	$\sum_{ij} p_i \cdot p_j$	(Gini)/2	balance-weighted variety
disparity	1	0	$\sum_{ij} d_{ij}$	(Solow & Polasky 1994*a*)	disparity-weighted variety
diversity	1	1	$\sum_{ij} d_{ij} \cdot p_i \cdot p_j$	D	balance/disparity-weighted variety

From (Napoli, 1999; Stirling, 2007)

(Semi-)formalising accountability

Principles for Accountable Algorithms

Responsibility

Make available externally visible avenues of redress for adverse individual or societal effects of an algorithmic decision system, and designate an internal role for the person who is responsible for the timely remedy of such issues.

Explainability

Ensure that algorithmic decisions as well as any data driving those decisions can be explained to end-users and other stakeholders in non-technical terms.

Accuracy

Identify, log, and articulate sources of error and uncertainty throughout the algorithm and its data sources so that expected and worst case implications can be understood and inform mitigation procedures.

Auditability

Enable interested third parties to probe, understand, and review the behavior of the algorithm through disclosure of information that enables monitoring, checking, or criticism, including through provision of detailed documentation, technically suitable APIs, and permissive terms of use.

Fairness

(Diakopoulos, Friedler, et al., 2016)

- These methods can be extremely useful for certain purposes.
- They also serve to raise awareness for a large number of side-effects of (semi-)automated decision-making.
- Still …

Problem solved?

OK. This was a rhetorical question.

5 Challenges

$= , \neq ?$

Challenge 1: Formalisation infiltrates the conception of human beings

(Berendt & Preibusch, 2014, 2017)

The importance of Aristotelean Equality in law

"The principle of equality requires that equal situations are treated equally and unequal situations differently. Failure to do so will amount to discrimination unless an objective and reasonable justification exists."

Explanatory memorandum protocol 12 to the ECHR

Equal treatment = non-discrimination?

“I treat all my employees the same –
They enter the office by the stairs.”



Taddeucci and McCall vs. Italy

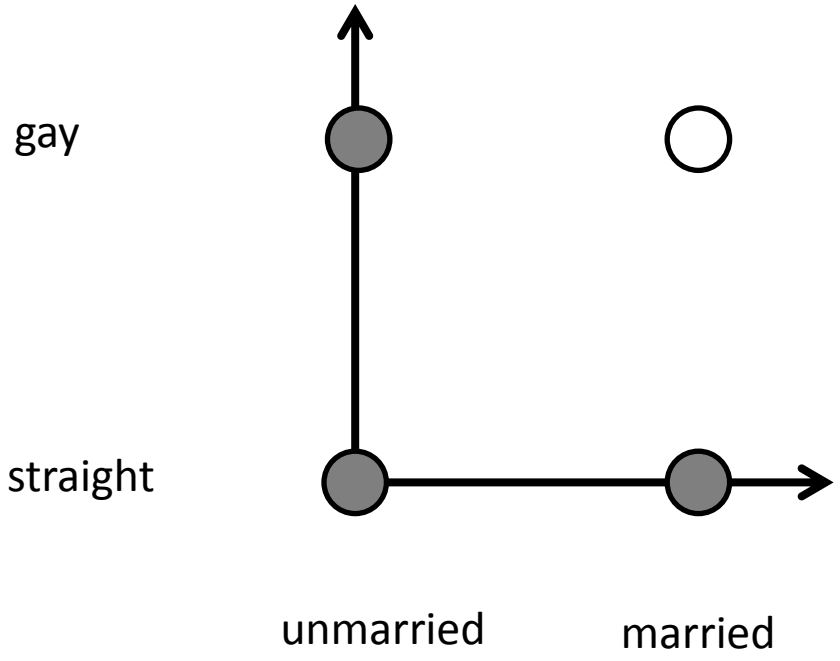
European Court of Human Rights judgment, 30/6/2016

The case concerned a refusal by the Italian authorities to grant a residence permit to a gay couple on family grounds.

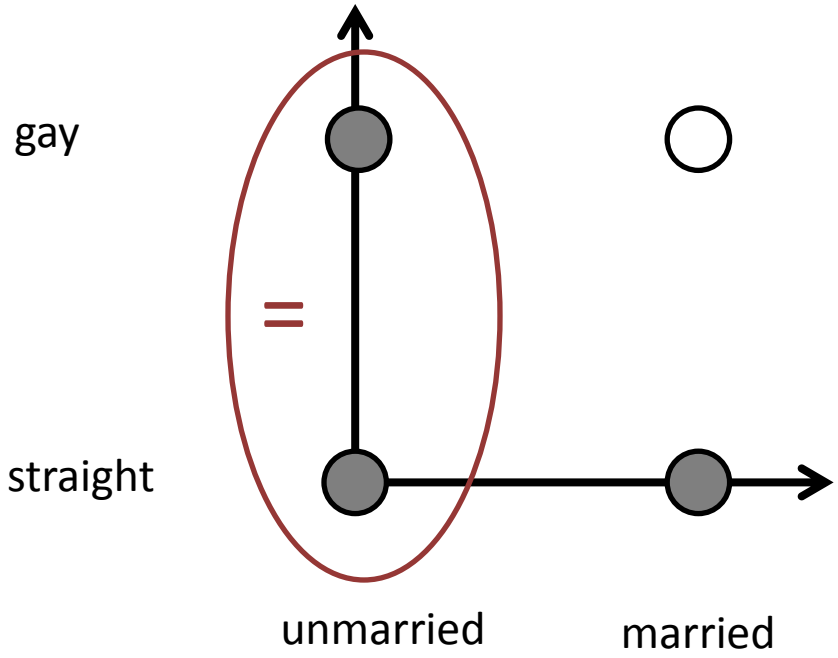
The Court found in particular that the situation of Mr Taddeucci and Mr McCall, a gay couple, could not be understood as comparable to that of an unmarried heterosexual couple. As they could not marry or, at the relevant time, obtain any other form of legal recognition of their situation in Italy, they could not be classified as “spouses” under national law. ....

Thus the Court concluded … that there had been a violation of Article 14 (prohibition of discrimination) taken together with Article 8 (right to respect for private and family life) of the European Convention on Human Rights.

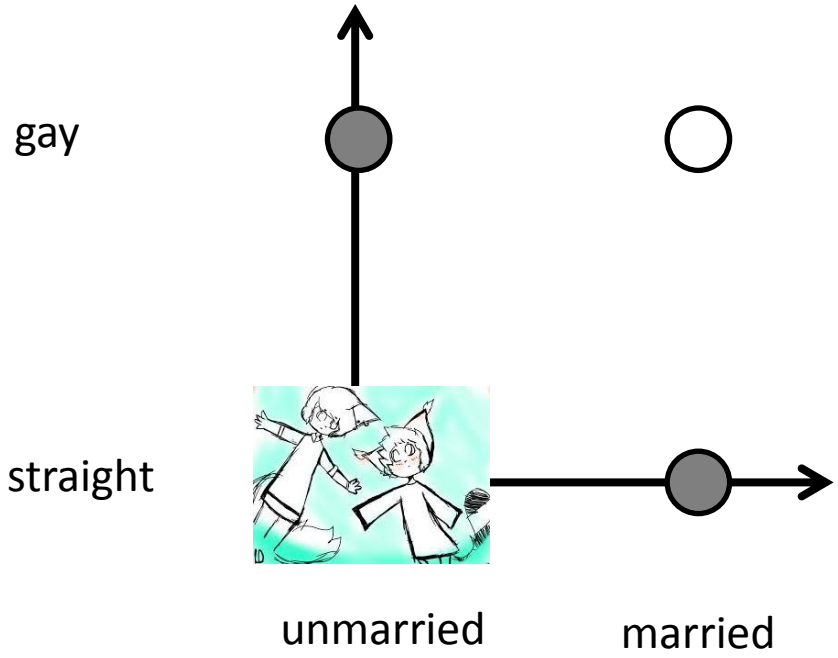
Taddeucci and McCall in feature space



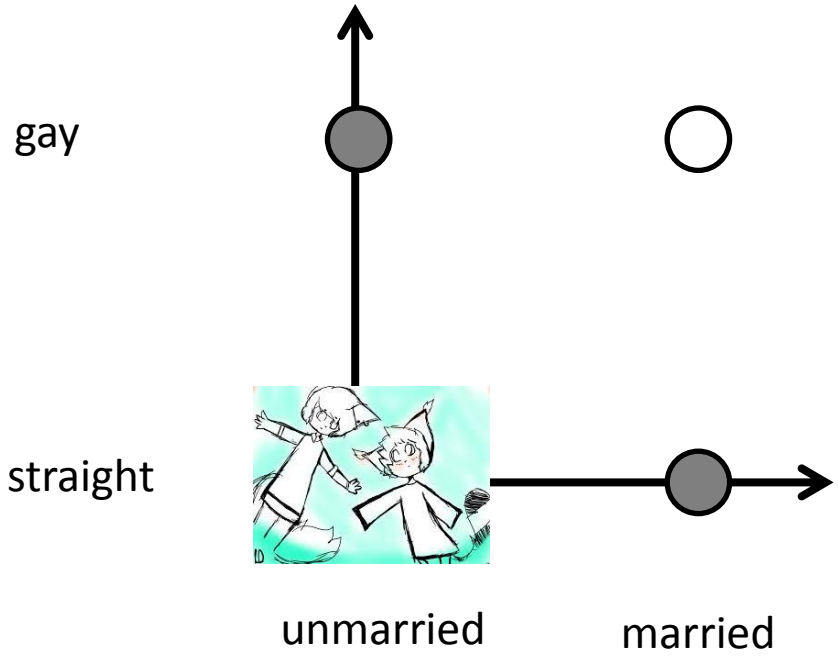
The Italian Government's opinion in feature space



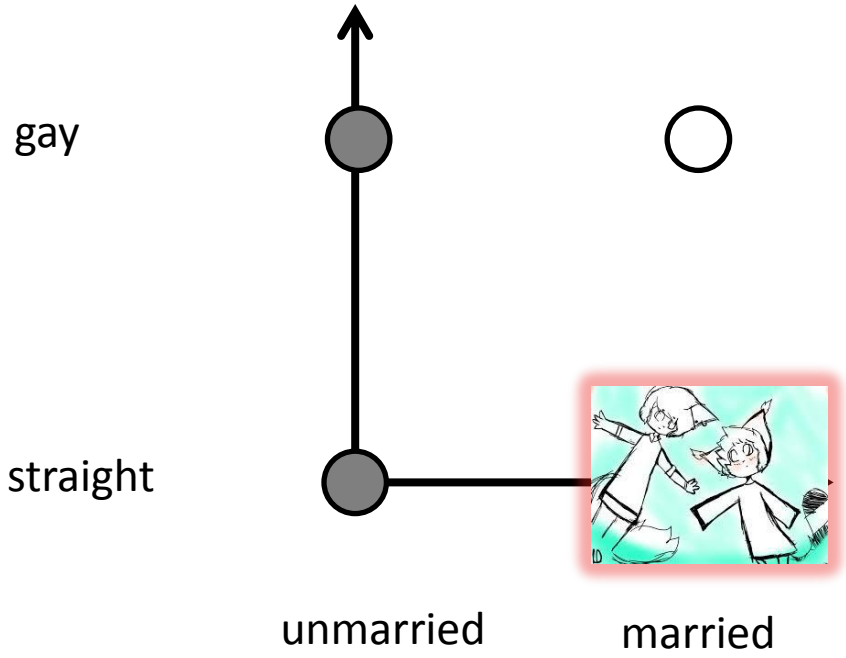
The ECHR's opinion in feature space



The ECHR's opinion in feature space



The ECHR's opinion in feature space



Human rights and their conception of human-ness

Equality, privacy, diversity as human rights, e.g.

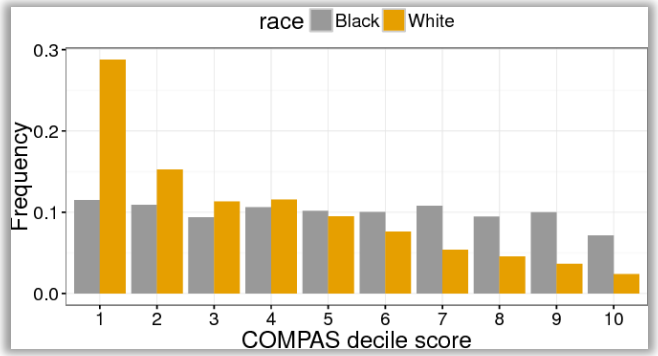
UDHR, Article 7: Right to Equality before the Law

Conception of human-ness in human rights?

"leading life in an autonomous, meaningful and responsible way"

(*"eigenständige, sinnhafte und verantwortliche Lebensführung*"; Brugger, *Das Menschenbild der Menschenrechte*, 1995)

ML methodology and conceptions of human-ness



Risk assessment and prediction in the criminal-justice sector

- *some scholars have framed the value-add of risk assessment strictly in terms of their ability to make predictions more accurately than the judges meting out such decisions by themselves …*
- *We argue for a shift away from predictive technologies, towards diagnostic methods that will help us to understand the criminogenic effects of the criminal justice system itself, as well as evaluate the effectiveness of interventions designed to interrupt cycles of crime.*
- *… these methods should be based in a more rigorous approach that incorporates both qualitative and quantitative data analysis.*

(Barabas et al., 2018)

Long-term unemployment / Câmara Municipal de Cascais, Portugal

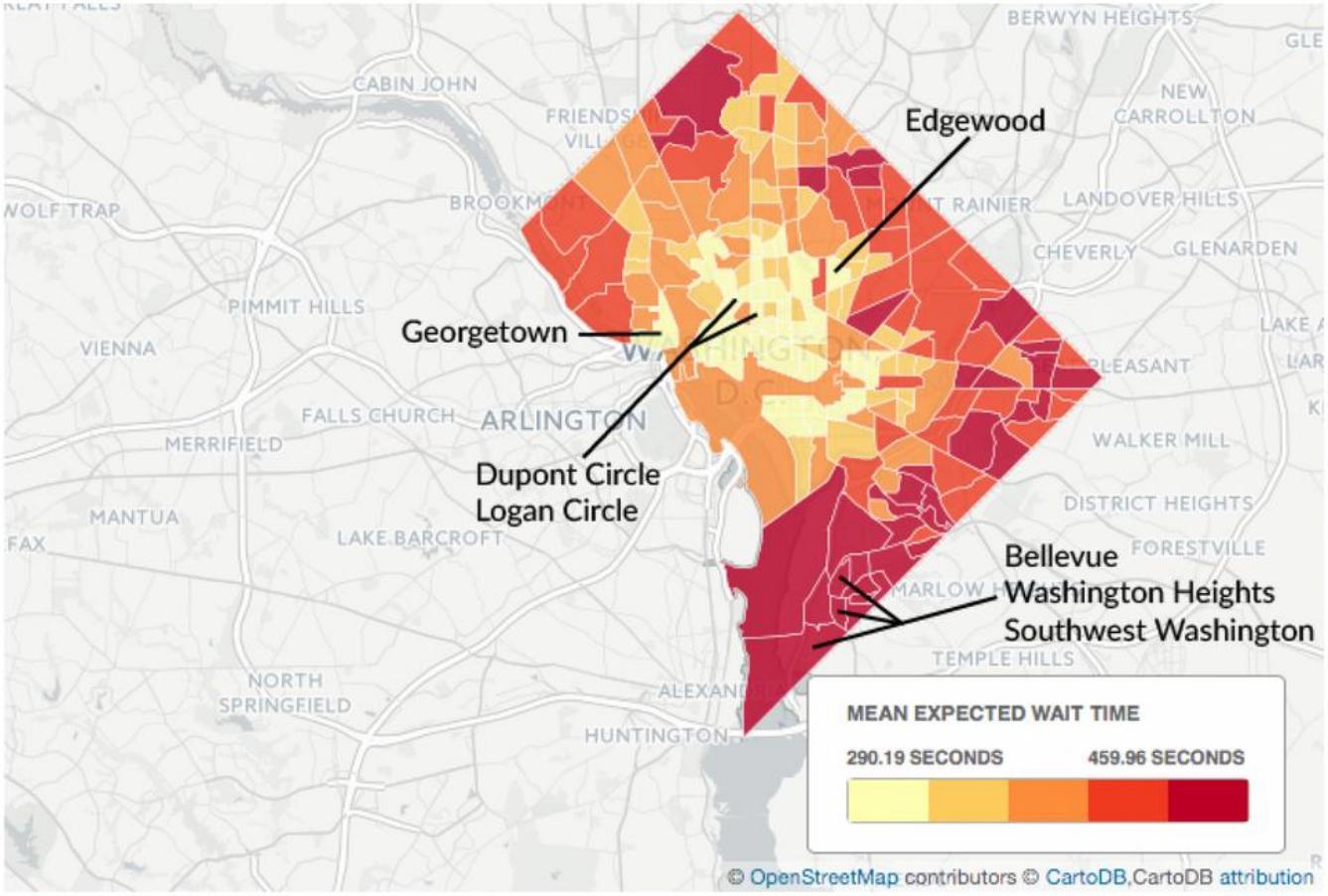
- One ex. of plans to build negative feedback loops

(Holmestad, 2017)

Who to protect?

Challenge 2: Formalisation may focus on some affected stakeholders (or their needs) at the expense of others

Transportation network companies: Discrimination?!



Census tracts with more people of color (including Black/African American, Asian, Hispanic-Black/African American, and Hispanic/Asian) have longer wait times. In other words, if you're in a neighborhood where there are more people of color, you'll wait longer for your uberX.

(Stark & Diakopoulos, 2016)

Transportation network companies: Other possible effects

- “Creative destruction”
- Integration of the sector
- Unfair competition with taxis and monopoly building



- Its cars or drivers may be unsafe or underinsured
- Invade customers' privacy
- Undermine working standards for taxi drivers and compensate its own drivers poorly

(Rogers, 2017)

The rating economy: emotional labour and disparate burden



- *Uber's rating system may require drivers, and perhaps even passengers, to engage in what has been called "emotional labor," or the work of establishing "micro-relationships that make customers feel good."*
- *To stay above a certain rating, drivers may need to be friendly, and perhaps a bit servile. Cab drivers, in comparison, can afford to be themselves-which may involve venting their frustration at long hours and low pay.*
- *Such emotional labor may impose a disparate burden on racial minorities. Minority drivers, to retain high ratings, may need to overcome white passengers' preconceptions, which can involve "identity work," or a conscious effort to track white, middle-class norms.*
- *Metaphorically, more and more workers [in the sharing economy] will be waiting tables.*

(Rogers, 2017)

… yet more stakeholders

Effect of transportation networks on the number of cars and amount of driving?

Cf.

parquetematización

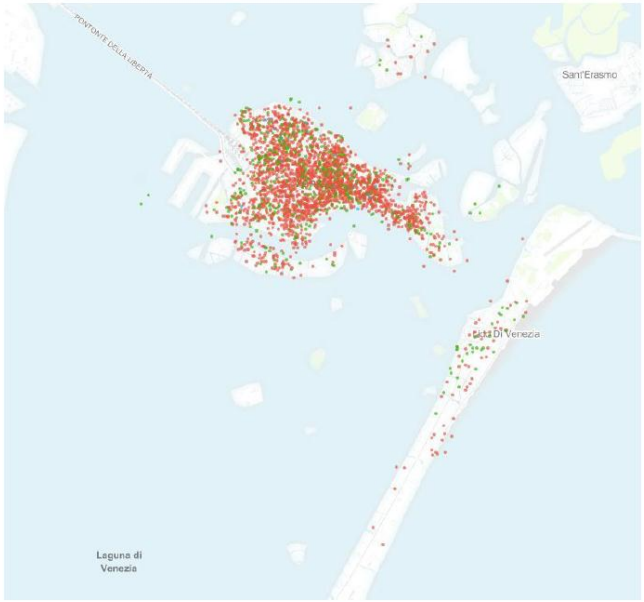


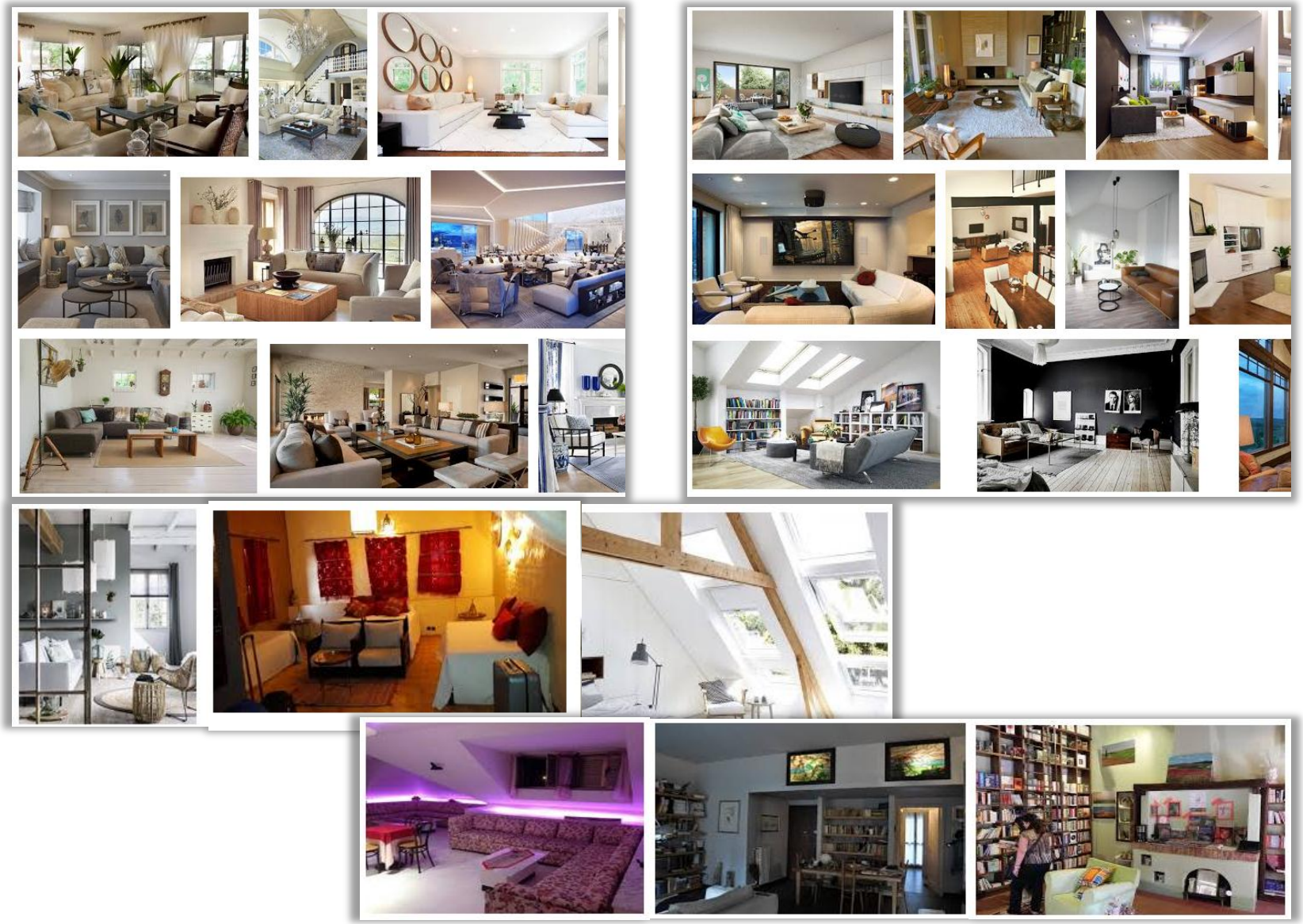
Figure 5: Image from InsideAirbnb (http://insideairbnb.com/venice/): Red dots represent full apartments or houses for rent, green dots represent single rooms for rent in a home.

Context vs. universality?

Challenge 3: Tools, terms, and what they make us forget

Transportation network companies in St. Petersburg

“beautiful living room” in 4 languages



Datasets as tools

Datasets/ AVA: A Large-Scale Database for Aesthetic Visual Analysis



Aesthetic Visual Analysis (AVA) contains over 250,000 images along with a rich variety of meta-data including a large number of aesthetic scores for each image, semantic labels for over 60 categories as well as labels related to photographic style for high-level image quality categorization.

dpchallenge
a digital photography contest

Become a Member Today!

Register for the site, and you can start participating right now! Submit to challenges, cast your votes, participate in the forums, and much more!

Register Now!

Image Search

Go

Provide Feedback

Have a suggestion? Found a bug? Please click here visit our site over GetSatisfaction and let your voice be heard!

Looking Up From Underneath - Open Challenge - Winning Entries (Sep. 12th - Sep. 18th)



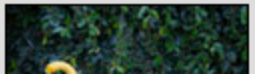
Migration Magic
nam



Alice, where did the rabbit go?
CNovack

A Vegetable Bestiary (DPL4 Final Challenge) - Member Challenge - Winning Entries (Sep. 10th - Sep. 16th)





Of whose aesthetics are these people (raters from dpchallenge) representative?

Name:
Username:
Type: Member

Gender: F
Location: Florida

Cameras: **Nikon D600**
Apple iPhone 6
Canon PowerShot SX100 IS
Nikon D80
Olympus C-4040Z

Lenses: **Nikon AF-S NIKKOR 24-70mm f/2.8G ED**
Nikon AF Nikkor 50mm f/1.4D
Nikon AF Zoom-Nikkor 18-135mm f/3.5- 5.6G ED DX
Nikon AF Zoom-Nikkor 70-200mm f/2.8 VR DG IF-ED
Nikon AF-S Micro-Nikkor 105mm f/2.8 VR

STUFF WHITE PEOPLE LIKE

This blog is devoted to stuff that white people like

Feeds: Posts Comments



Does a social value in itself incentivise decision makers?

Challenge 4: Formalisation oversimplifies the agency of (decision making) stakeholders

DIAMOND project
e.g. (Peperkamp & Berendt, 2018)

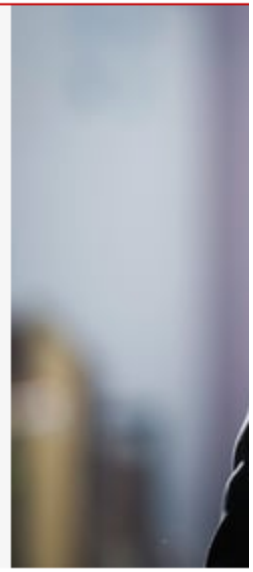
Interfaces that don’t work so well (1): Diversity in news consumption

Burst your bubble

The Guardian's weekly guide to conservative articles worth reading to expand your thinking

31 May 2018

Dinesh D'Souza: far-right provocateur and key figure in US culture wars

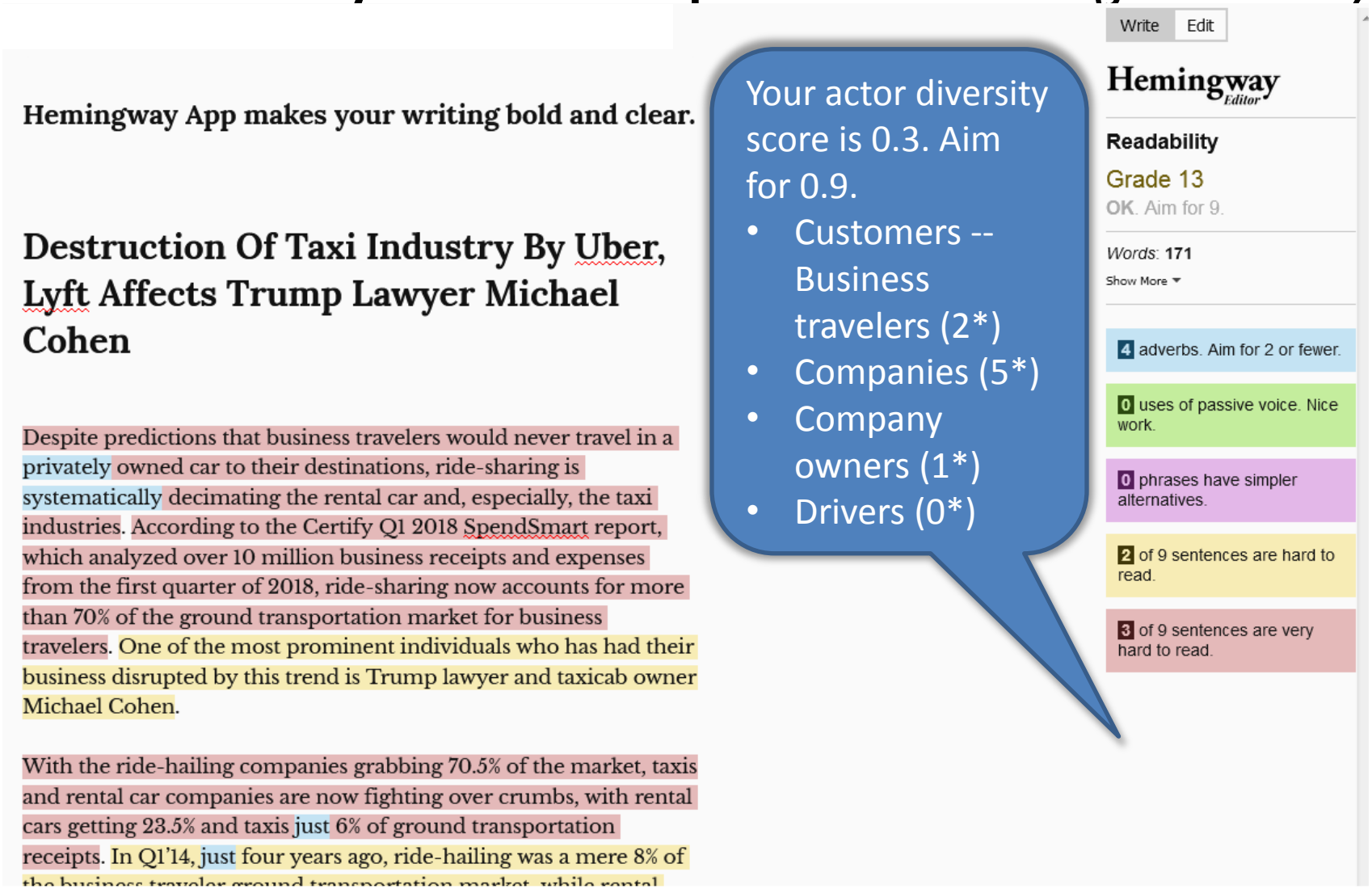


Conservative commentator pardoned by Trump is a big fan of the president and has special enmity for Clinton and Obama

6:48 PM

NB: This slide should not be misconstrued – The Guardian does a great job here!

Interfaces that don't work so well (2): Diversity in news production *(fictitious)*



Building bridges between a rights view and a technical view

Challenge 5: Overreliance on formal concepts neglects essential but non-cognitive drivers of behaviour

(Schiffner, Berendt, et al., 2018; Work to come in VeriLearn)

What to do if you entrust some agent with a relevant task?

When it's possible to prove correct execution:

- Verification

When this is not possible:

- Accountability

Accountability: political science

(Bovens, 2007)

A relationship qualifies as a case of accountability when:

1. There is a relationship between an actor and a forum
2. in which the actor is obliged
3. to explain and justify
4. his conduct,
5. the forum can pose questions,
6. pass judgement,
7. and the actor may face consequences.

Accountability: political science

(Bovens, 2007)

A relationship qualifies as a case of accountability when:

1. There is a relationship between an actor and a forum
2. in which the actor is obliged
3. to explain and justify
4. his conduct,
5. the forum can pose questions,
6. pass judgement,
7. and the actor may face consequences.

The GDPR: a law towards accountable data protection

A relationship qualifies as a case of accountability when:

1. There is a relationship between an actor and a forum
2. in which the actor is obliged
3. to explain and justify
4. his conduct,
5. the forum can pose questions,
6. pass judgement,
7. and the actor may face consequences.

"Technical and organisational measures" relating to

processing of personal data per se

GDPR obligations

Rec. 71, Art. 13-15, 20

Art. 5 (2)

Art. 13-15, 20, 22; 5(2)

) Art.

) 83, 58

Building bridges: The GDPR and its "state of the art" requirements

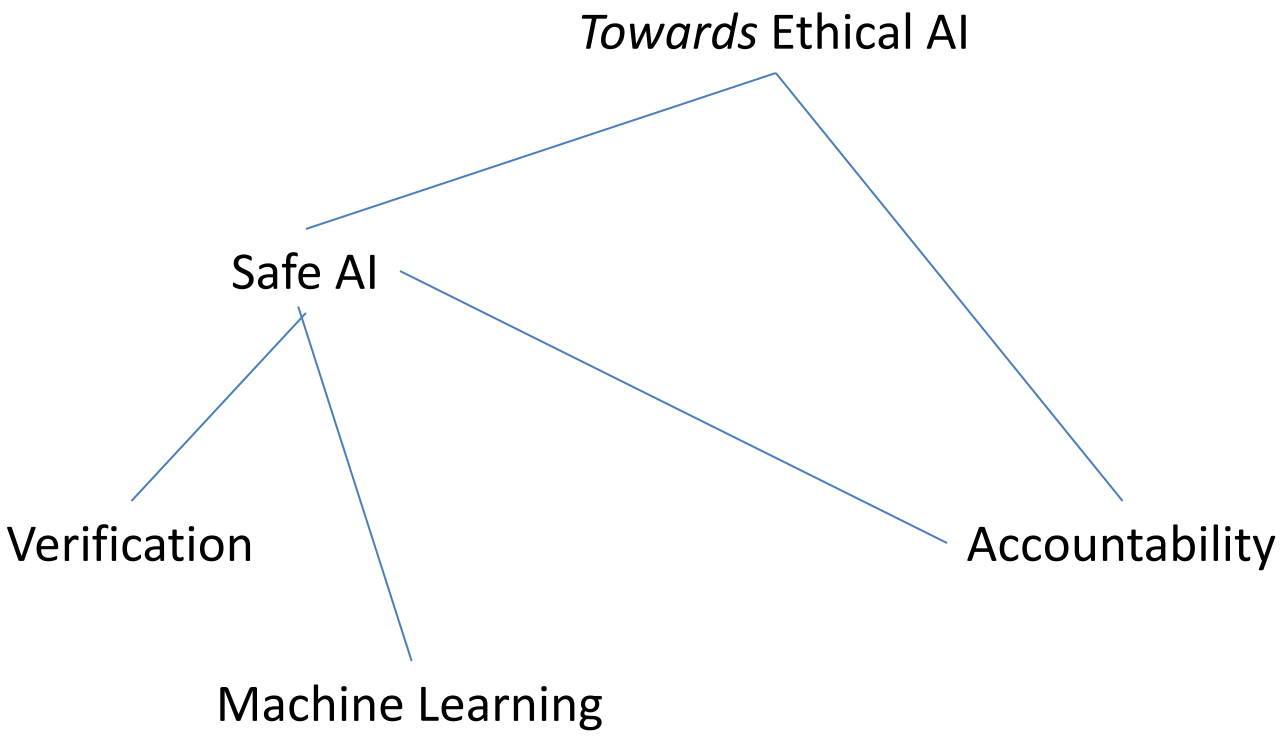
- Article 24(1) requires controllers to implement appropriate technical and organisational measures to ensure and to be able to demonstrate that processing of personal data is performed in accordance with the GDPR.
- Articles 25(1) and 32(1) require controllers and processors to give due regard to the state of the art when choosing the technologies.

A CS task! E.g. ENISA PETs repository

- available technologies?
- burden their adoption adds for controller and processor?
- actual functionality (protection goals)?
- relation to the legal requirements?

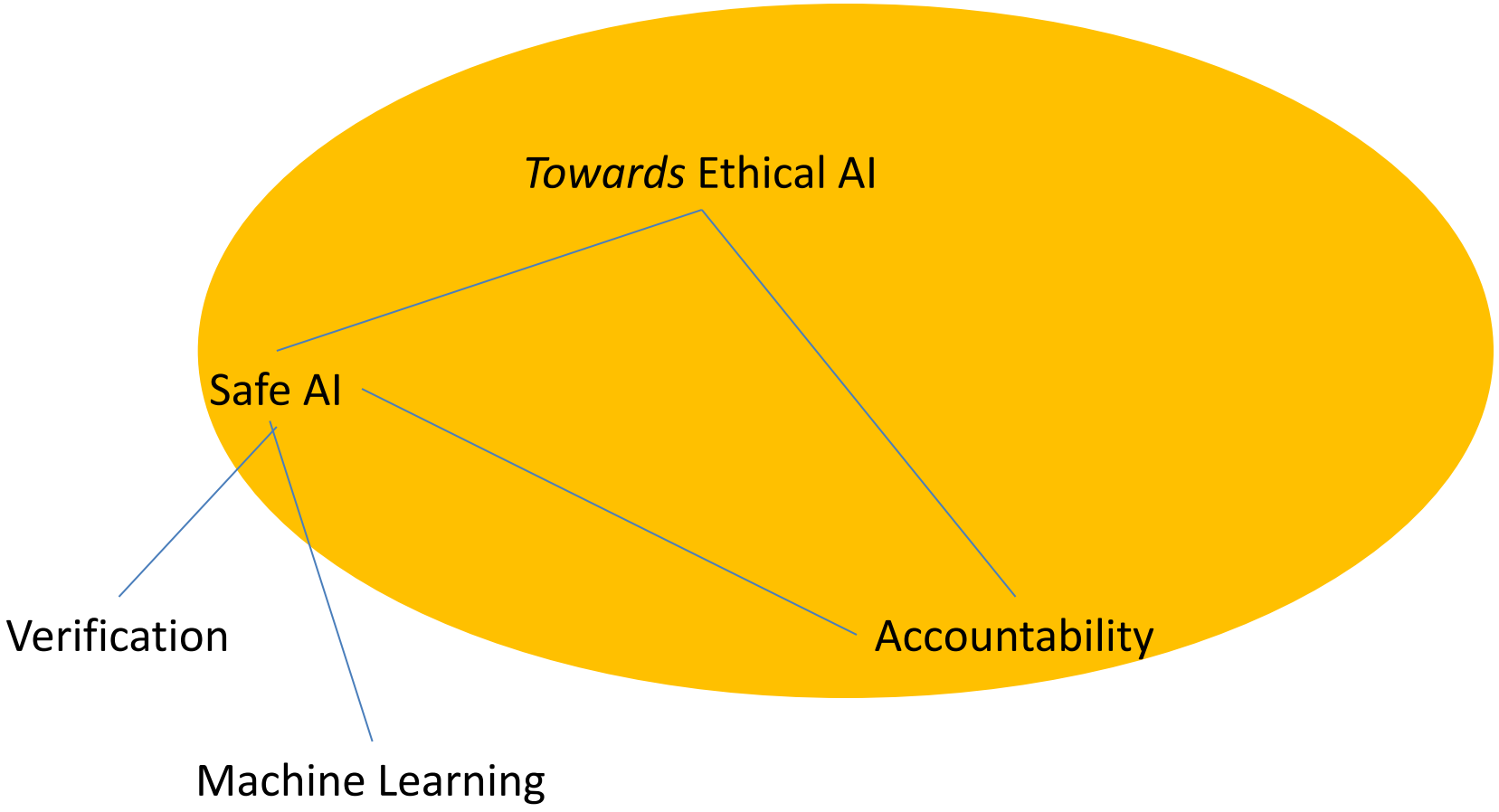
The VeriLearn project

https://dtai.cs.kuleuven.be/projects/verilearn



The VeriLearn project

https://dtai.cs.kuleuven.be/projects/verilearn



Thank you!

References

C. Barabas, K. Dinakar, J. Ito, M. Virza, J. Zittrain (2018). Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment. Proceedings of Machine Learning Research, 81, http://proceedings.mlr.press/v81/barabas18a/barabas18a.pdf

Berendt, B. & Preibusch, S. (2014). Better decision support through exploratory discrimination-aware data mining: foundations and empirical evidence. Artificial Intelligence and Law, 22 (2), 175-209. https://people.cs.kuleuven.be/~bettina.berendt/Papers/berendt_preibusch_2014.pdf

Berendt, B. & Preibusch, S. (2017). Toward accountable discrimination-aware data mining: The importance of keeping the human in the loop – and under the looking-glass. Big Data, 5(2). DOI: 10.1089/big.2016.0055.
https://people.cs.kuleuven.be/~bettina.berendt/Papers/berendt_preibusch_2017_last_author_version.pdf

Bovens M. (2007). Analysing and assessing public accountability. A conceptual framework. European Law Journal, 13, 447–468.

Brugger, W. (1995). Das Menschenbild der Menschenrechte. In Jahrbuch für Recht und Ethik / Annual Review of Law and Ethics, Vol. 3, Themenschwerpunkt: Rechtsstaat und Menschenrechte / Human Rights and the Rule of Law (1995), pp. 121-134.

Diakopoulos, N., Friedler, S., et al. (2016). Principles for Accountable Algorithms and a Social Impact Statement for Algorithms
http://www.fatml.org/resources/principles-for-accountable-algorithms

Holmestad et al. (2017). Predicting Risk of Long-term Unemployment. Presentation at the Data Science for Social Good Conference 2017, Sep. 28-29, 2017, Chicago, IL. https://dssg.uchicago.edu/wp-content/uploads/2017/09/holmestad.pdf

Napoli, P.M. (1999). Deconstructing the diversity principle. Journal of Communication, 49(4), 7-34.

Jeroen Peperkamp, Bettina Berendt (2018). Diversity Checker: Toward Recommendations for Improving Journalism with Respect to Diversity. UMAP (Adjunct Publication) 2018: 35-41.

Pedreschi D, Ruggieri S, Turini F (2008) Discrimination-aware data mining. In: Proceedings of KDD'08, pp 560–568. ACM.

Dino Pedreschi, Salvatore Ruggieri, Franco Turini (2009). Integrating induction and deduction for finding evidence of discrimination. ICAIL 2009: 157-166.

Rogers, B. (2017). The Social Costs of Uber. University of Chicago Law Review Online, Vol. 82 [2017], Iss. 1, Art. 6, 85-102.

Schiffner, S., Berendt, B., Siil, T., Degeling, M., Riemann, R., Schaub, F., Wuyts, K., Attoresi, M., Gürses, S., Klabunde, A., Polonetsky, J., Sadeh, N., Zanfir-Fortuna, G. (2018). Towards a roadmap for privacy technologies and the General Data Protection Regulation: A transatlantic initiative. To appear in Proceedings of the Annual Privacy Forum 2018, 13-14 June 2018, Barcelona, Springer LNCS.
https://people.cs.kuleuven.be/~bettina.berendt/Papers/schiffner_et_al_APF_2018.pdf

Stark, J. & Diakopoulos, N. (2016). Uber seems to offer better service in areas with more white people. That raises some tough questions. The Washington Post Wonkblog, March 10, 2016. https://www.washingtonpost.com/news/wonk/wp/2016/03/10/uber-seems-to-offer-better-service-in-areas-with-more-white-people-that-raises-some-tough-questions/

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. J. R. Soc. Interface 4,

Sweeney, L. (2013). Discrimination in Online Ad Delivery. Communications of the ACM, 56(5), 44-54. cited from earlier version available at SSRN:

http://ssrn.com/abstract=2208240 or http://dx.doi.org/10.2139/ssrn.2208240

Binh Luong Thanh, Salvatore Ruggieri, Franco Turini (2011). k-nn as an implementation of situation testing for discrimination discovery and prevention. In: KDD, pp 502–510. ACM.