On limits of AI transparency –
thoughts from a computational perspective

Bettina Berendt
KU Leuven

Slides including references at
https://people.cs.kuleuven.be/~bettina.berendt/

FRA Panel *What does transparency mean for detecting and avoiding discriminatory AI-based decisions?*
at the Amsterdam Privacy Conference 2018. 7 Oct. 2018.

How to operationalise transparency?




IL GELATO
È PIÙ BUONO
SE SAI COME
È FATTO
OPERAZIONE
TRASPARENZA
LIBRO DEGLI INGREDIENTI:
TU SAI COSA MANGI
E NOI SIAMO
ORGOGLIOSI
DI SPIEGARTI IL
NOSTRO PRODOTTO.
GELATIERIPERILGELATO.COM
GELATIERI
PER IL GELATO
PASSIONE
ARTIGIANALE





"How the ~~ice cream~~ AI decision was made"
The 1970s+: traces (1)

…

It has already been established that

[1.1] the gram stain of ORGANISM-1 is gramneg, and

[1.2] the morphology of ORGANISM-1 is rod, and

[1.3.] the aerobicity of ORGANISM-1 is facultative

…

****HOW 1.3?**

[i.e. HOW was it established that the aerobicity of ORGANISM-1 is facultative?]

The following were used:

[3.1] RULE027 indicated there is strongly suggestive evidence (.8) that the aerobicity of ORGANISM-1 is facultative

[2.1] RULE027 indicated there is weakly suggestive evidence (.2) that the aerobicity of ORGANISM-1 is anaerobic

Since this gave a cumulative CF of (.8) for facultative, and (.2) for anaerobic, it has been established that the aerobicity of ORGANISM-1 is facultative.

“How the AI decision was made”
The 1970s+: traces (2)

…

It has already been established that

[1.1] the gram stain of ORGANISM-1 is gramneg, and

[1.2] the morphology of ORGANISM-1 is rod, and

[1.3.] the aerobicity of ORGANISM-1 is facultative

…

**** HOW 1.1?**

[i.e., HOW was it established that the gram stain of ORGANISM-1 is gramneg?]

You said so [question 11]

“How the AI decision was made” The 1970s+: traces + strategy (3)

15) Is the patient's illness with ORGANISM-1 a hospital acquired infection?

****WHY?**

[i.e., WHY is it important to determine whether or not the infection with ORGANISM-1 was acquired while the patient was hospitalized?]

[1.0] This will aid in determining the category of ORGANISM-1.

It has already been established that

[1.1] the gram stain of ORGANISM-1 is gramneg, and

[1.2] the morphology of ORGANISM-1 is rod, and

[1.3.] the aerobicity of ORGANISM-1 is facultative

Therefore, if

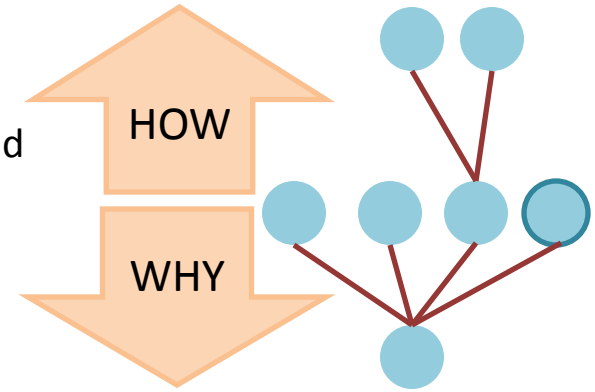
[1.4] the infection with ORGANISM-1

was not acquired while the patient was hospitalized

then there is strongly suggestive evidence (.8)

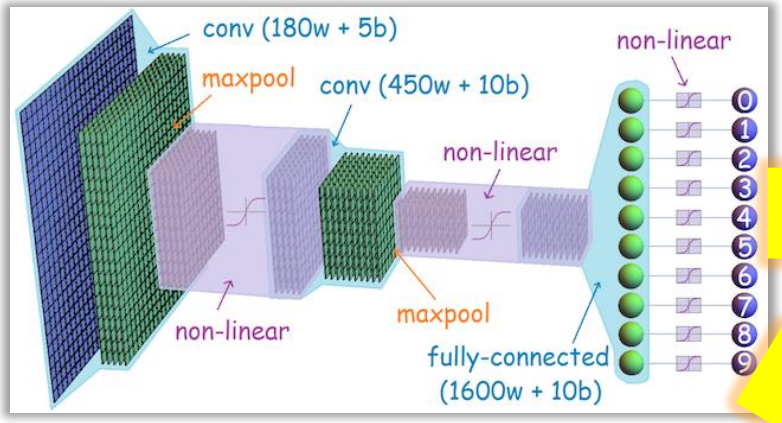
that the category of ORGANISM-1 is enterobacteriaceae

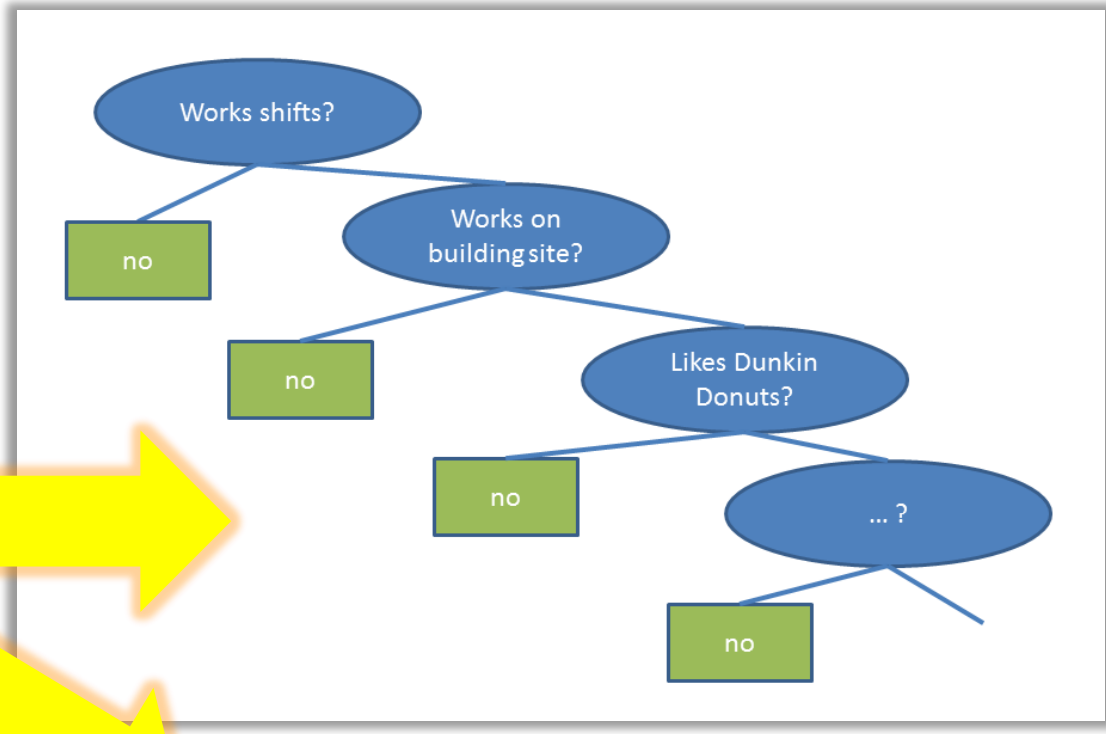
ALSO: there is weakly suggestive evidence (.1) that the identity ORGANISM-1 is pseudomonas-aeruginosa [RULE037].

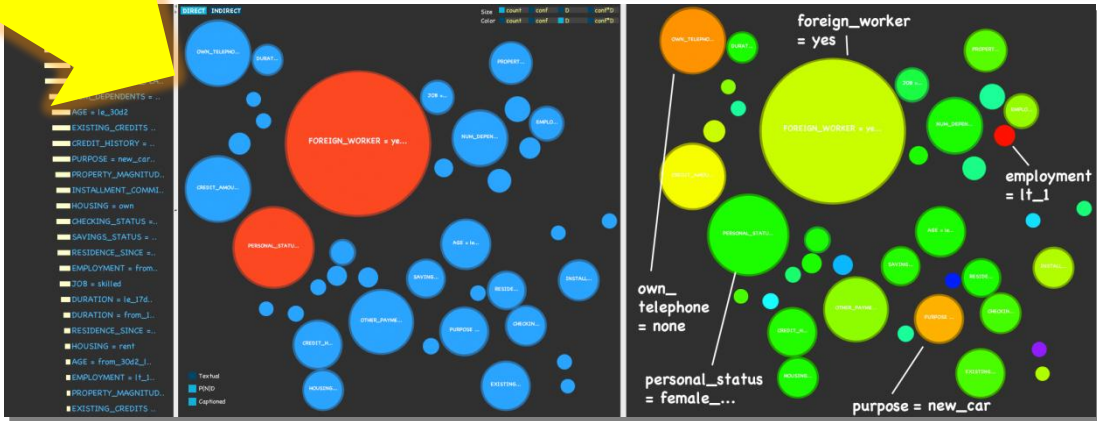


AI/ML explanations today *(very simplified)*:

"creating understandable model", "highlighting relevant detail"







The FRA steps towards building fair algorithms

- checking the quality of the data being used to build algorithms to avoid faulty algorithm ‘training’;
- promoting transparency – being open about the data and code used to build the algorithm, as well as the logic underlying the algorithm, and providing meaningful explanations of how it is being used. Among others, this will help individuals looking to challenge data-based decisions pursue their claims;
- carrying out impact assessments that focus on the implications for fundamental rights, including whether they may discriminate based on protected grounds, and seeing how proxy information can produce biased results;
- involving experts in oversight: to be effective, reviews need to involve statisticians, lawyers, social scientists, computer scientists, mathematicians and experts in the subject at issue.

Data quality

Detect, Prevent

Explain

Challenge

FR impact assessm.

Auditors

Interdisciplinary

What AI/ML algorithmic explanations & algorithmic modifications can do

- Detect biases (along attribute lines) in data
- Prevent biases/unfairness in decisions,
- Explain descriptively how the decision was made / what the model distinguishes
- Support auditing (e.g., logs)

If the grounds are well-defined, and
the notion of fairness matches the meta-model of machine learning

What AI/ML algorithmic explanations & algorithmic modifications cannot do

- Tell whether biases in data are a data quality issue (could be facts)
- Detect or prevent biases/unfairness with a more complex / FR-based meta-model (equality principle)
- Give normative justifications (e.g., choice of fairness measure)
- Provide a causal explanation when the model is correlative
- By themselves provide for possibility to challenge decisions
- Take a systemic perspective of the systems around the algorithm (concepts, implementation, data, people)
  - This also includes possible auditors!
  - a general (though not definitional) problem: scarcity of theory and evidence base that the explanations are meaningful

How do you challenge a fact?

Or: just one ex. of the importance of a system view

- Correlative / actuarial
  - Poor people cannot repay loans.
    - e.g., microcredit, 1976/1983+
  - Women live longer.
    - ECJ Test-Achats judgement, 2011
  - People who buy in guitar shops have bad loan-repayment behaviour. (Lori Andrews, 2012)
    - ?
- Causal (?)
  - Men take more risks because they have more testosterone.
    - Cordelia Fine, *Testosterone Rex*, 2017

What transparency via AI/ML algorithmic explanations may also do

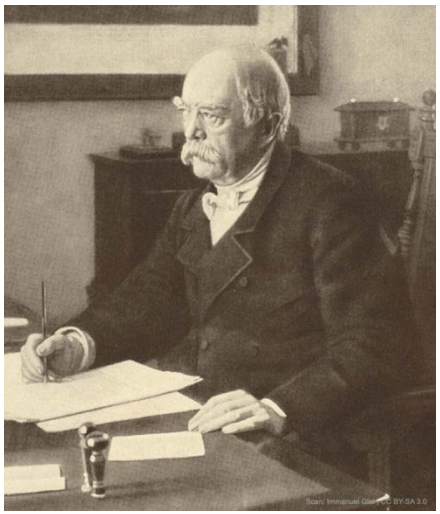
- Violate privacy
  - reveal data/information about individuals
  - … not to mention about groups …
- Violate secrecy, e.g. for security
  - allow for reverse engineering of the algorithm / method
- Allow people to game the system
  - Whether that's bad or good depends on your perspective!

Countermeasures

- Intervenability
- Embed algorithmic approaches in an organisational accountability structure, following design patterns (→ data protection by design)
- Apply data privacy methods to explanations
  - up to: “Accountability without transparency”

- ”The less people know about how sausages and laws are made, the better they sleep.“







A better metaphor for algorithm-related transparency



A better metaphor for algorithm-related transparency



A better metaphor for algorithm-related transparency



Systemic

- for individuals
- within society
- water AND sanitation

Needs education & participation;

Not all details needed/desired



Is a human right

Requires auditors & interdisciplinarity

Some reading (1)

(All sources are available online.)

1970s explanations: The MYCIN system was presented in

Shortliffe, E.H. & Buchanan, B.G. (1975). A model of inexact reasoning in medicine. Mathematical Biosciences, 23(3-4), 351-379.

Overviews of, and reflections on, machine learning explanation approaches

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Pedreschi, D., & Giannotti, F. (2018). A Survey Of Methods For Explaining Black Box Models. arxiv:1802.01933v3

Lipton, Z.C. (2016). The Mythos of Model Interpretability. ICML 2016 Workshop on Human Interpretabilityin Machine Learning (WHI 2016).

including a visualisation of the discriminatory power of individual features:

Berendt, B. & Preibusch, S. (2012). Exploring discrimination: A user-centric evaluation of discrimination-aware data mining. In 2012 IEEE 12th Intern. Conference on Data Mining Workshops (pp. 344-351). IEEE Computer Science Press. https://people.cs.kuleuven.be/~bettina.berendt/Papers/berendt_preibusch_dpadm2012_with_publication_note.pdf

FRA white paper on algorithmic discrimination and fairness

FRA (European Union Agency for Fundamental Rights) (2018). #BigData: Discrimination in data-supported decision making.

Literature overviews of discrimination detection and prevention

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 . http://people.cs.kuleuven.be/~bettina.berendt/Papers/berendt_preibusch_2014.pdf

More recent overviews: see various tutorials by, among other people, Sara Hajian or Solon Barocas

Explanation and justification: inspiration taken from

Strandburg, K. (2016). Decisionmaking, Machine Learning and the Value of Explanation. Presentation at the Workshop The Human Use of Machine Learning: An Interdisciplinary Workshop, Dec. 16, 2016, Venice, Italy.

Ye, R.L. (1995). The value of explanation in expert systems for auditing: An experimental investigation. Expert Systems with Applications, 9(4), 543-556.

Some reading (2)

Sources cited concerning "facts"

Andrews, L. (2012). I Know Who You Are and I Saw What You Did: Social Networks and the Death of Privacy. Simon and Schuster.

Fine, C. (2017). Testosterone Rex: Unmaking the Myths of Our Gendered Minds. Icon.

Transparency in conflict with various aspects of privacy:

Atzori, M., Bonchi, F., Giannotti, F., & Pedreschi, D. (2008). Anonymity preserving pattern discovery. The VLDB Journal, 17(4), 703-727. [and many works since]

Hansen, M., Jensen, M., & Rost, M. (2015). Protection goals for privacy engineering. Presentation at International Workshop on Privacy Engineering, 21 May 2015.

Transparency in the context of the whole GDPR and privacy engineering context:

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. In Proccedings of the Annual Privacy Forum 2018, 13-14 June 2018, Barcelona. Springer.

Intervenability

Hansen, M. (2015). The art of intervenability for privacy engineering. Presentation at the Workshop Data Protection, Privacy, and Transparency (DPPT'15), Hamburg, 26 May 2015

Herrmann, T., Schmidt, A., & Degeling, M. (2018). From interaction to intervention: An approach for keeping humans in control in the context of socio-technical systems. In Proceeding of STPIS'18.

Accountability by design

Berendt, B. & Preibusch, S. (2016). Towards accountable non-discriminatory data mining: The importance of keeping the human in the loop – and under the looking-glass. Big Data, 5(2), 135-152.

“Accountability without transparency”

Kroll, J.A., Huey, J., Barocas, S., Felten, E.W., Reidenberg, J.E., Robinson, D.G., & Yu, H. (2017). Accountable algorithms. University of Pennsylvania Law Review, 165.