Explainable AI, Fairness and Bias

Jelke Bloem

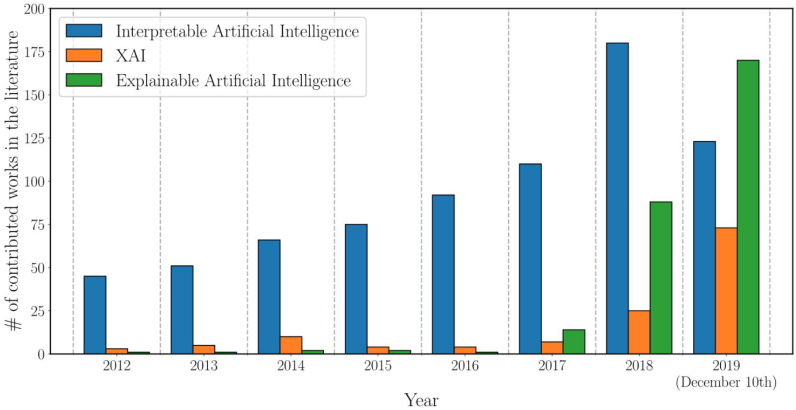
Text Mining
Amsterdam University College
With materials from AI BSc course AI for Society

May 9, 2025

Announcements

- Friday 09/05: Reading Assignment 5: Bias in word embeddings
- Friday 09/05: Project update 2

A recent trend...



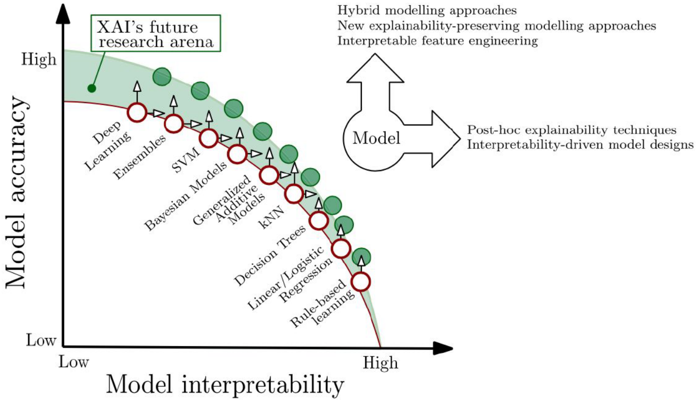
`https://arxiv.org/abs/1910.10045`

Explainable and Interpretable AI

- The definitions are not very clear yet, as it is an emerging field
- Interpretation: how does a model work? (model transparency)
  - Allow human to grasp the mechanism used to come up with a decision
- Explanation: what can a model tell me? (post-hoc reasoning)
  - Deconstruct steps that were used in making a decision

Explain to whom?

Performance vs Interpretability tradeoff



Social aspects of the explanation/interpretation

- Confidence: grows when the rationale of a decision is close to the thought processes of the user
- Trust: grows when decisions do not require validation to be acted upon
- Safety: the system is consistent and relible, displays uncertainty or confidence level, is robust to outliers etc.
- Ethics: the system does not violate a certain well-defined code of principles

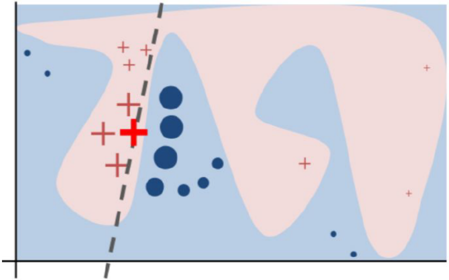
`https://arxiv.org/abs/2004.14545`

Contextual aspects of the explanation/interpretation

- Contrastive: identify elements unique to this decision
- Selective: provide the most relevant causes
- Provide causes: humans are bad at interpreting probabilities
- Social context: may call for different kind of explanation

LIME: Local Interpretable Model-agnostic Explanations

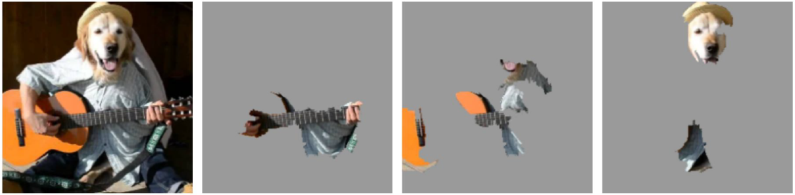
- Algorithm that explains predictions of a classifier by approximating it locally (in the vicinity of the predicted data point) with an interpretable model
- Treat original model as black box
- Train simple interpretable linear classifier on input features and classification decision



https://arxiv.org/abs/1602.04938

LIME: Example

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \; \mathcal{L}(f, g, \pi_x) + \Omega(g)$$



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

+ SP-LIME: Method to select representative examples of a classification problem to show to the user

`https://arxiv.org/abs/1602.04938`

Explainable AI

- Can we have explanation without interpretability?
- Can people accurately explain how they make decisions?

Links on Explainable AI

- List of libraries to explain black-box models: https://github.com/EthicalML/awesome-production-machine-learning#explaining-black-box-models-and-datasets
- LIME implementation in Python: https://github.com/marcotcr/lime
- SHAP unifies LIME and many more methods: https://github.com/slundberg/shap
- AIX360: https://github.com/Trusted-AI/AIX360
- Language Interpretability Tool (from UvA): https://github.com/pair-code/lit

Understanding Language Models: BERTology

- How do you study a black box language model?

`https://arxiv.org/pdf/2002.12327.pdf`

BERTology questions

- Does BERT base itself on the syntax of human language or just on the linear order of the words?
- Is syntactic structure in the attention weights or in the token representations?
- Does BERT understand negation?
- Does BERT know subject-verb agreement?
- Does BERT understand numbers?
- BERT as a knowledge base?

BERTology methods

- Probing classifiers
  - Use hidden states or attention weights as input to a classifier that predicts a linguistic property of the input text
- Visualization
- Input perturbation
- Masked Language Modeling task
- Nonce word task
- Model perplexity/surprisal

Masked Language Modeling example

AllenNLP Interpret
https://allennlp.org/interpret

Ai2 Allen Institute for AI — AllenNLP

Simple Gradients Visualization

See saliency map interpretations generated by visualizing the gradient.

Saliency Map:

[CLS] The [MASK] rushed to the emergency room to see her patient . [SEP]

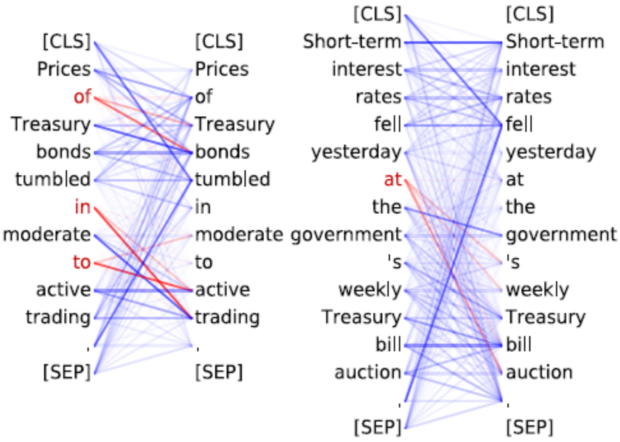
Mask 1 Predictions:

- 47.1% nurse
- 16.4% woman
- 10.0% doctor
- 3.4% mother
- 3.0% girl

Visualization example

Head 9-6

- **Prepositions** attend to their objects
- 76.3% accuracy at the pobj relation



https://arxiv.org/pdf/1906.04341.pdf

Knowledge Base example

AtLocation	You are likely to find a overflow in a ____.	drain	sewer [-3.1] , canal [-3.2] , toilet [-3.3] , stream [-3.6] , **drain [-3.6]**
CapableOf	Ravens can ____.	fly	**fly [-1.5]** , fight [-1.8] , kill [-2.2] , die [-3.2] , hunt [-3.4]
CausesDesire	Joke would make you want to ____.	laugh	cry [-1.7] , die [-1.7] , **laugh [-2.0]** , vomit [-2.6] , scream [-2.6]
Causes	Sometimes virus causes ____.	infection	disease [-1.2] , cancer [-2.0] , **infection [-2.6]** , plague [-3.3] , fever [-3.4]
HasA	Birds have ____.	feathers	wings [-1.8] , nests [-3.1] , **feathers [-3.2]** , died [-3.7] , eggs [-3.9]
HasPrerequisite	Typing requires ____.	speed	patience [-3.5] , precision [-3.6] , registration [-3.8] , accuracy [-4.0] , **speed [-4.1]**
HasProperty	Time is ____.	finite	short [-1.7] , passing [-1.8] , precious [-2.9] , irrelevant [-3.2] , gone [-4.0]
MotivatedByGoal	You would celebrate because you are ____.	alive	happy [-2.4] , human [-3.3] , **alive [-3.3]** , young [-3.6] , free [-3.9]
ReceivesAction	Skills can be ____.	taught	acquired [-2.5] , useful [-2.5] , learned [-2.8] , combined [-3.9] , varied [-3.9]
UsedFor	A pond is for ____.	fish	swimming [-1.3] , fishing [-1.4] , bathing [-2.0] , **fish [-2.8]** , recreation [-3.1]

`https://aclanthology.org/D19-1250.pdf`

Links on Explainable AI

- List of libraries to explain black-box models: https://github.com/EthicalML/awesome-production-machine-learning#explaining-black-box-models-and-datasets
- LIME implementation in Python: https://github.com/marcotcr/lime
- SHAP unifies LIME and many more methods: https://github.com/slundberg/shap
- AIX360: https://github.com/Trusted-AI/AIX360
- Language Interpretability Tool (from UvA): https://github.com/pair-code/lit

Fairness and bias in AI

Fairness in AI

- Not a very clearly defined concept
- Lack of bias in decisions
- Balanced treatment of sub-populations and individuals
- Equality of opportunity
- Equity in outcomes

Definition of fairness are often mutually exclusive (mathematically and morally).

Some attempts at formal definitions of fairness in AI:
https://arxiv.org/abs/1901.10002
https://www.annualreviews.org/doi/abs/10.1146/annurev-statistics-042720-125902
https://arxiv.org/abs/1908.09635

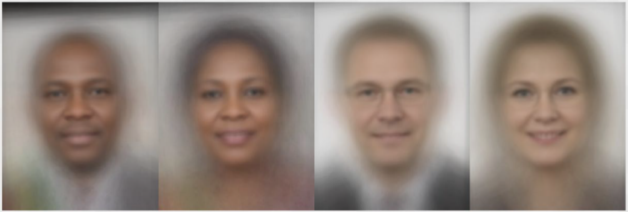
Types of definitions

- **Group-Independent Predictions** require that the decisions that are made are independent (or conditionally independent) of group membership. For example, the demographic parity criterion states that the proportion of each segment of a protected class (e.g., gender) should receive the positive outcome at equal rates.
- **Equal Metrics Across Groups** require equal prediction metrics of some sort (this could be accuracy, true positive rates, false positive rates, and so on) across groups. For example, the equality of opportunity criterion requires equal true positive/negative rates across groups.
- **Individual Fairness** requires that individuals who are similar with respect to the prediction task are treated similarly. There is an assumption that an ideal feature space exists in which to compute similarity, and that those features are recoverable in the available data. For example, fairness through (un)awareness tries to identify a task-specific similarity metric in which individuals who are close according to this metric are also close in outcome space.
- **Causal Fairness** definitions place some requirement on the causal graph that generated the data and outcome. For example, counterfactual fairness requires that there is not a causal pathway from a sensitive attribute to the outcome decision

`https://arxiv.org/abs/1901.10002`

Bias in facial recognition

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

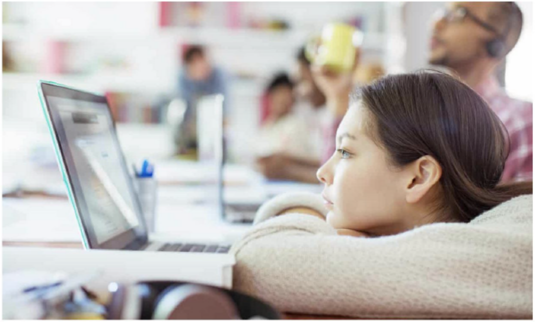


`http://gendershades.org/index.html`

Bias in job ad recommendation

Women less likely to be shown ads for high-paid jobs on Google, study shows

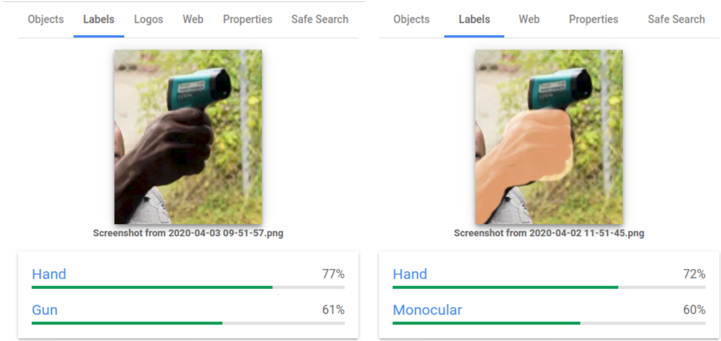
Automated testing and analysis of company's advertising system reveals male job seekers are shown far more adverts for high-paying executive jobs



▲ One experiment showed that Google displayed adverts for a career coaching service for executive jobs 1,852 times to the male group and only 318 times to the female group. Photograph: Alamy

https://www.theguardian.com/technology/2015/jul/08/women-less-likely-ads-high-paid-jobs-google-study

Bias in Google Vision AI

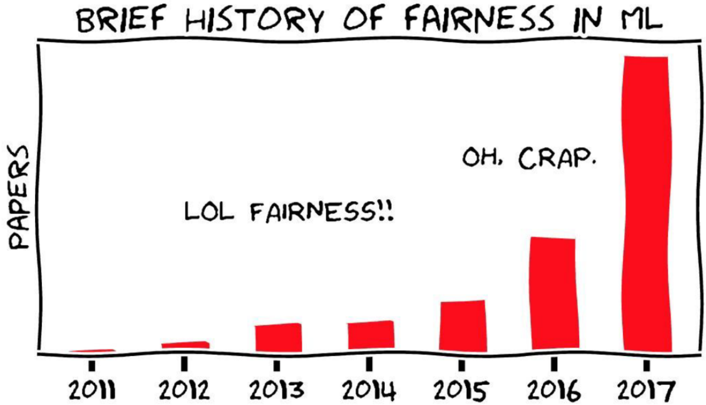


https://algorithmwatch.org/en/google-vision-racism/

Consequences of lack of fairness

- **Impact of error types**: Sometimes a false positive (being falsely recognized as a shoplifter) is worse than a false negative (being falsely flagged as innocent)
- **Disparate impact**: Being flagged as holding a gun by error usually has worse consequences than being flagged holding something else by error.
- **Allocative harm**: Unfair allocation of resources (e.g. hiring decisions)
- **Representational harm**: Unfair depiction of individuals or groups (e.g. stereotyping)

Kate Crawford's lecture 'The trouble with bias': `https://www.youtube.com/watch?v=fMym_BKWQzk`

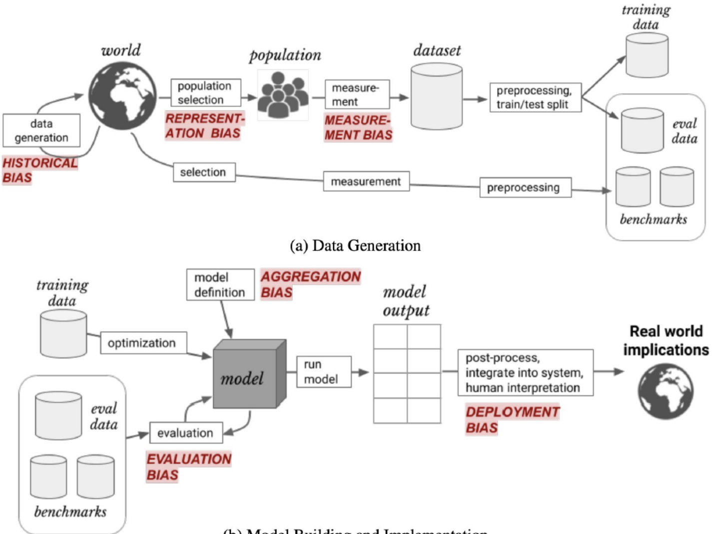


Credit: Moritz Hardt

Fairness

- Who is responsible for algorithmic unfairness?

Bias in AI



(a) Data Generation

(b) Model Building and Implementation

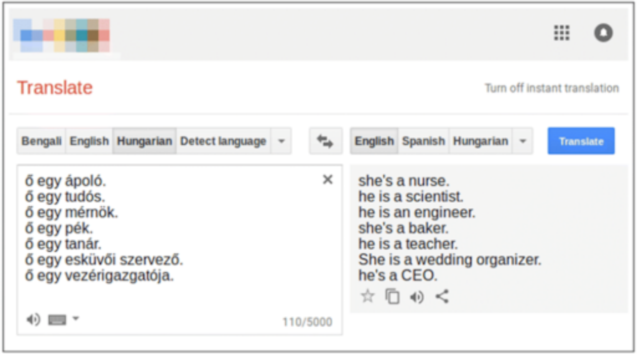
Bias

- Stereotypical bias
- Statistical bias
- Cognitive bias
  - https://upload.wikimedia.org/wikipedia/commons/a/a4/The_Cognitive_Bias_Codex_-_180%2B_biases%2C_designed_by_John_Manoogian_III_%28jm3%29.png

Considering bias in building AI systems

- Define what bias and fairness means in the context of your task
- Explore your data: skewness, outliers, missing values, unbalance across protected groups. Avoid possible bias in data acquisition
- Consider underrepresented and protected groups in model evaluation.
- Consider intersections of protected/underrepresented groups
- Consider possibly unintended consequences when deploying
- Ask for diverse feedback (especially from protected groups involved)

Gender bias and stereotyping



Measuring (gender) bias in word embeddings

- Define a set of "definitional word pairs" that capture the gender dimension (e.g., he/she, man/woman, etc.)
- Measure bias by how differently a word *w* projects onto word pairs.
  - $x\_he = cos(\text{"politician"}, \text{"he"})$
  - $x\_she = cos(\text{"politician"}, \text{"she"})$
  - x_he - x_she = measure of bias towards the masculine gender

Bolukbasi et al. (2016): `https://arxiv.org/abs/1607.06520`

Measuring (gender) bias in word embeddings

Identify the gender subspace:

- Consider the pairwise differences among the set of "definitional word pairs" that capture the gender dimension (he-she, etc.)
- Apply dimensionality reduction on them (e.g. PCA), and find the gender subspace.
- Use the cosine between any word and this gender subspace to quantify its bias. This bias can be averaged over a set of words.
- If you take masculine - feminine, a positive cosine might be indicative of bias towards the masculine gender, vice versa for a negative one.

Dealing with (gender) bias in word embeddings

- Neutralize and equalize (**hard de-biasing**): enforces that any gender neutral word is set to zero onto the gender subspace.
- Soften (**soft de-biasing**): Ensures that neutral words are equidistant from equality sets. For example, it ensures that brother, sister and husband, wife are both equidistant from babysitting, although probably the latter set will still be closer than the former.

Approaches to bias in word embeddings

1. Work on data (e.g. filtering the training corpus)
2. Work on the algorithm (loss, bias mitigation via a constrained optimization objective)
3. Post-hoc methods (transforming the embeddings in some way)

`https://www.aclweb.org/anthology/P19-1159`
`https://www.aaai.org/AAAI22Papers/AISI-6900.DingL.pdf`