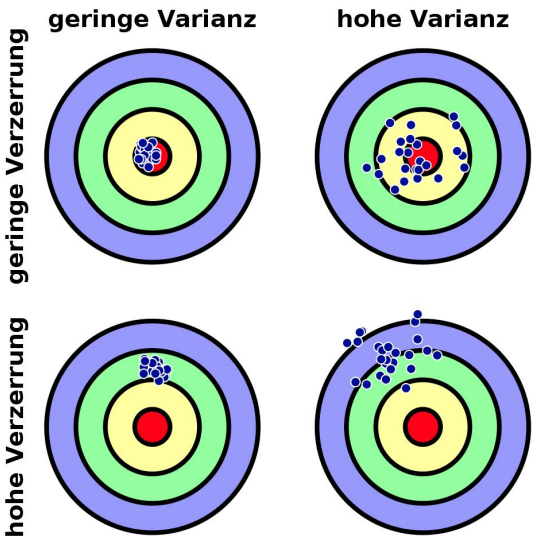
Bias in LMs

What is Bias?

- dt. “Verzerrung”
- “The bias error is an error from erroneous assumptions in the learning algorithm. High bias can cause an algorithm to miss the relevant relations between features and target outputs” (Wikipedia: Bias-Variance Tradeoff)



Source: https://de.wikipedia.org/wiki/Verzerrung-Varianz-Dilemma#/media/Datei:Varianz_und_Bias.svg

Bias in NLP. Case Study: Gender

Sentence

the programmer finished [MASK] work

Mask 1

Prediction	Score
the programmer finished **his** work	54.4%
the programmer finished **the** work	38.3%
the programmer finished **its** work	2%
the programmer finished **her** work	1.3%
the programmer finished **their** work	0.7%

Sentence

the nurse finished [MASK] work

Mask 1

Prediction	Score
the nurse finished **her** work	77.4%
the nurse finished **the** work	14.4%
the nurse finished **his** work	4.7%
the nurse finished **its** work	0.9%
the nurse finished **to** work	0.5%

AllenNLP

Source: “Ethics and Bias”: https://drive.google.com/drive/folders/1nvfoSi9enjf6IroUWsAz39ocoe2MJKKf

the (programmer/nurse) finished [MASK] work

- probably, most programmers in the world are male, and most nurses are female (check the stats if you like)
- This fact is represented in LM training data
- So maybe this behavior is fine, right?

So maybe some bias is okay?

- Amazon built a tool to rank applications
- The tool was trained on real applications and rankings
- The tool discriminated against women in technical jobs
- That’s a problem
  - legally
  - for product builders
  - for society

https://towardsdatascience.com › ... ▾ Diese Seite übersetzen

Sorry Ladies… This Is Still a MAN's World! | by Jillian Chambers

von J Chambers — So, by now you must have heard and perhaps forgotten about **Amazon's** automated **hiring** tool that discriminates based on **gender** (Dastin, 2018).

https://www.aclu.org › blog › why... ▾ Diese Seite übersetzen

Why Amazon's Automated Hiring Tool Discriminated ...

12.10.2018 — And it's not just **gender discrimination** we should be concerned about. Think about all the ways in which looking at resume features might ...

https://www.businessinsider.com › ... ▾ Diese Seite übersetzen

No Surprise Amazon's AI Was Biased Against Women ...

13.10.2018 — **Amazon** abandoned a project to build a machine learning program for **recruitment** which engineers found was discriminating against female ...

https://www.bbc.com › news › tech... ▾ Diese Seite übersetzen

Amazon scrapped 'sexist AI' tool - BBC News

10.10.2018 — ... **recruitment** tool was abandoned after it showed **bias** towards men. ... that the system was not rating candidates in a **gender**-neutral way ...

https://theconversation.com › amaz... ▾ Diese Seite übersetzen

Amazon's sexist hiring algorithm could still be better than a ...

01.11.2018 — Although **Amazon** is at the forefront of AI technology, the company couldn't find a way to make its algorithm **gender**-neutral.

https://mashable.com › article › am... ▾ Diese Seite übersetzen

Amazon's sexist recruiting algorithm reflects a larger ...

10.10.2018 — **Amazon** developed a **recruiting** algorithm that would rank candidates, but it devalued women, and shows sexism in the workplace.

Reasons for hiring bot misbehavior

- Penalized terms:
  - activities like *women's chess club* etc.
  - all-women college names
- Rewarded terms:
  - Terms that are more frequent in male CVs: *captured, executed*
- Problem: Hiring is intransparent. Not clear to candidate why CV is rejected
- Conclusion: Maybe one day, a hiring algorithm is fairer than human recruiters. But that requires hard work
- General Question: How to separate world knowledge (frequent genders of jobs) and stereotypes?