

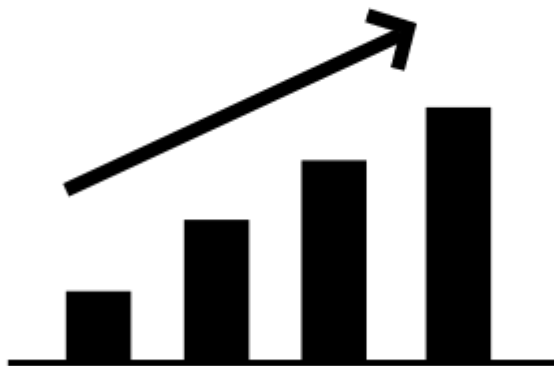
Responsible Text Mining

Dong Nguyen
29th of July, 2024



Utrecht University

Advances in NLP

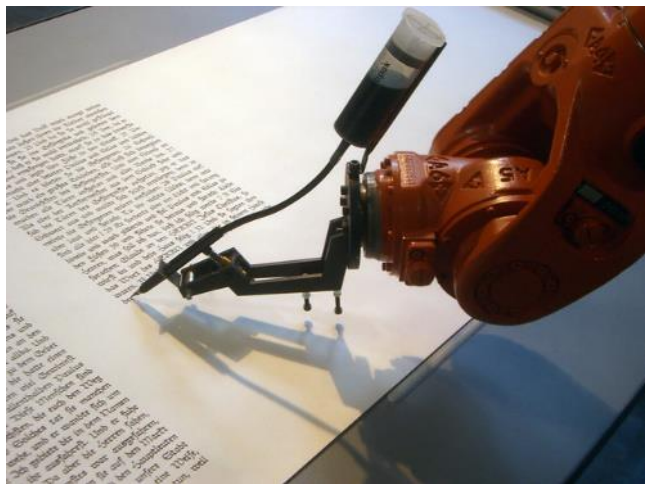


performance



Dual Use

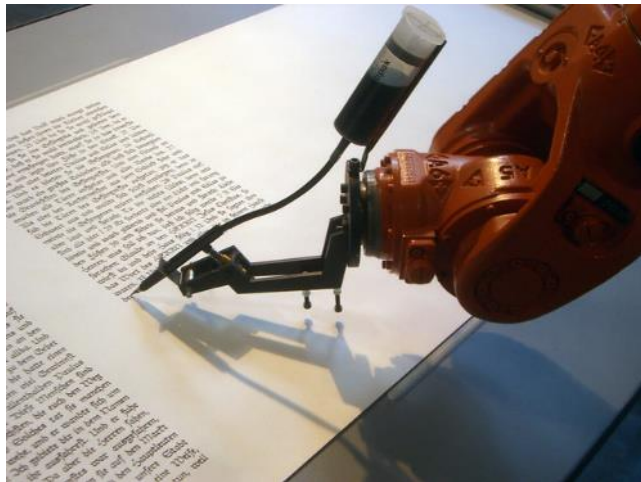
Dual use: Text generation



Generate novels,
poems, summaries

AI systems might be used for both beneficial and harmful purposes

Dual use: Text generation



Generate novels,
poems, summaries

Disinformation



AI systems might be used for both beneficial and harmful purposes

Dual use: Should I build this system?

Predicting Depression via Social Media

Munmun De Choudhury

Michael Gamon

Scott Counts

Eric Horvitz

Microsoft Research, Redmond WA 98052

{munmund, mgamon, counts, horvitz}@microsoft.com

“We explore the potential to use social media to detect and diagnose major depressive disorder in individuals.”

How can such a system be used
for a beneficial purpose?
How can such a system be used
for a harmful purpose?

(3 min)

Are we really
measuring what we
intend to measure?



What can go
wrong?



7 x 2

Are horses clever?

If the eighth day of the month
comes on a Tuesday, what is the
date of the following Friday?

Clever Hans

Claimed to have
performed
arithmetic and
other intellectual
tasks.



Wolf or dog?



Can the system really distinguish between dogs and wolves?



Sentiment analysis



8/10

Sci-fi perfection. A truly mesmerizing film.

I'm nearly at a loss for words. Just when you thought Christopher Nolan couldn't follow up to "The Dark Knight", he does it again, delivering another masterpiece, one with so much power and rich themes that has been lost from the box office for several years. Questioning illusions vs reality usually makes the film weird, but Nolan grips your attention like an iron claw that you just can't help watching and wondering what will happen next. That is a real powerful skill a director has. No wonder Warner Bros. put their trust in him, he is THAT good of a director, and over-hyping a Christopher Nolan film, no matter what the film is about, is always an understatement instead of an overestimate like MANY films before.

Models can be right for the wrong reasons 😞

Is our model actually measuring what we think it is measuring?

Next

*Behavioral
testing of NLP
models*

Explainability

Next

*Behavioral
testing of NLP
models*

Explainability

Behavioral testing of (black-box) NLP models



That cabin crew is extraordinary

Sentiment analysis.
This text is? positive, negative, neutral

*Beyond Accuracy: Behavioral Testing of NLP Models with
CheckList, Ribeiro, Wu, Guestrin and Singh, at ACL 2020 [\[link\]](#)*

Behavioral testing of (black-box) NLP models



That cabin crew is extraordinary

Sentiment analysis.
This text is? **positive**, negative, neutral

*Beyond Accuracy: Behavioral Testing of NLP Models with
CheckList, Ribeiro, Wu, Guestrin and Singh, at ACL 2020 [\[link\]](#)*

Behavioral testing of (black-box) NLP models



That cabin crew is extraordinary

Sentiment analysis:
{**positive**, negative, neutral}

Beyond Accuracy: Behavioral Testing of NLP Models with CheckList, Ribeiro, Wu, Guestrin and Singh, at ACL 2020 [\[link\]](#)

CheckList:

- Switching person names shouldn't change predictions
 - **Sharon** -> **Erin** was great (inv)
- Author sentiment is more important than of others
 - *Some people hate you, but I think you are exceptional* (pos)
- etc...

Behavioral testing of NLP models: Hatecheck

Automatic detection of hate speech is incredibly difficult

The New York Times

THE FACEBOOK PAPERS

In India, Facebook Grapples With an Amplified Version of Its Problems

Internal documents show a struggle with misinformation, hate speech and celebrations of violence in the country, the company's biggest market.

<https://www.nytimes.com/2021/10/23/technology/facebook-india-misinformation.html> (Oct 23, 2021)

Meta Platforms Inc. took action in the third quarter against more than 28 million pieces of content on Facebook and Instagram that violated its policies against hate speech. The vast majority of hateful posts that required action were on Facebook, the company said, noting that prevalence of hate speech is well under 1% for both social media platforms.

<https://www.bloomberg.com/news/articles/2021-11-09/facebook-takes-action-on-hate-speech-amid-whistle-blower-claims> (Nov 9, 2021)

HateCheck: Creating the test suite

Initial list

- Review of previous research
- Interviews with 21 employees from NGOs who work with online hate

people will be quoting someone, calling that person out [...] but that will get picked up by the system

I have seen the algorithm get it wrong, if someone's saying something like 'I'm so gay'.

A lot of LGBT people use slurs to identify themselves, like reclaim the word queer, and people [...] report that and then that will get hidden

Behavioral testing of NLP models: HateCheck

Functionality	Example Test Case	Gold Label	accuracy		
			B-D	B-F	P
Expression of strong negative emotions (explicit)	I hate [IDENTITY].	hateful	88.6	90.7	98.6
Reclaimed slurs	Shout out to all my n*ggas.	non-hate	39.5	33.3	28.4
Non-hate expressed using negated hateful statement	No [IDENTITY] deserves to die	non-hate	12.8	12.0	3.8
Added spaces between chars	I want to k i l l all [IDENTITY].	hateful	43.9	37.6	74.0
Denouncements of hate that make direct reference to it	You have to stop calling [IDENTITY] disgusting.	non-hate	29.1	29.8	18.4

B-D: BERT fine-tuned on Davidson et al. (2017)
B-F: BERT fine-tuned on Founta et al. (2018)
P: Perspective

Röttger, Vidgen, Nguyen, Waseem, Margetts and Pierrehumbert. HateCheck: Functional tests for hate speech detection models. ACL 2021

Behavioral testing of NLP models: HateCheck

Functionality	Example Test Case	Gold Label	accuracy		
			B-D	B-F	P
Expression of strong negative emotions (explicit)	I hate [IDENTITY].	hateful	88.6	90.7	98.6
Reclaimed slurs	Shout out to all my n*ggas.	non-hate	39.5	33.3	28.4
Non-hate expressed using negated hateful statement	No [IDENTITY] deserves to die	non-hate	12.8	12.0	3.8
Added spaces between chars	I want to k i l l all [IDENTITY].	hateful	43.9	37.6	74.0
Denouncements of hate that make direct reference to it	You have to stop calling [IDENTITY] disgusting.	non-hate	29.1	29.8	18.4

B-D: BERT fine-tuned on Davidson et al. (2017)
B-F: BERT fine-tuned on Founta et al. (2018)
P: Perspective

Röttger, Vidgen, Nguyen, Waseem, Margetts and Pierrehumbert. HateCheck: Functional tests for hate speech detection models. ACL 2021

Behavioral testing of NLP models: HateCheck

Functionality	Example Test Case	Gold Label	accuracy		
			B-D	B-F	P
Expression of strong negative emotions (explicit)	I hate [IDENTITY].	hateful	88.6	90.7	98.6
Reclaimed slurs	Shout out to all my n*ggas.	non-hate	39.5	33.3	28.4
Non-hate expressed using negated hateful statement	No [IDENTITY] deserves to die	non-hate	12.8	12.0	3.8
Added spaces between chars	I want to k i l l all [IDENTITY].	hateful	43.9	37.6	74.0
Denouncements of hate that make direct reference to it	You have to stop calling [IDENTITY] disgusting.	non-hate	29.1	29.8	18.4

B-D: BERT fine-tuned on Davidson et al. (2017)
B-F: BERT fine-tuned on Founta et al. (2018)
P: Perspective

Röttger, Vidgen, Nguyen, Waseem, Margetts and Pierrehumbert. HateCheck: Functional tests for hate speech detection models. ACL 2021

Behavioral testing of NLP models: HateCheck

Functionality	Example Test Case	Gold Label	accuracy		
			B-D	B-F	P
Expression of strong negative emotions (explicit)	I hate [IDENTITY].	hateful	88.6	90.7	98.6
Reclaimed slurs	Shout out to all my n*ggas.	non-hate	39.5	33.3	28.4
Non-hate expressed using negated hateful statement	No [IDENTITY] deserves to die	non-hate	12.8	12.0	3.8
Added spaces between chars	I want to k i l l all [IDENTITY].	hateful	43.9	37.6	74.0
Denouncements of hate that make direct reference to it	You have to stop calling [IDENTITY] disgusting.	non-hate	29.1	29.8	18.4

B-D: BERT fine-tuned on Davidson et al. (2017)
B-F: BERT fine-tuned on Founta et al. (2018)
P: Perspective

Röttger, Vidgen, Nguyen, Waseem, Margetts and Pierrehumbert. HateCheck: Functional tests for hate speech detection models. ACL 2021

Behavioral testing of NLP models: HateCheck

Functionality	Example Test Case	Gold Label	accuracy		
			B-D	B-F	P
Expression of strong negative emotions (explicit)	I hate [IDENTITY].	hateful	88.6	90.7	98.6
Reclaimed slurs	Shout out to all my n*ggas.	non-hate	39.5	33.3	28.4
Non-hate expressed using negated hateful statement	No [IDENTITY] deserves to die	non-hate	12.8	12.0	3.8
Added spaces between chars	I want to k i l l all [IDENTITY].	hateful	43.9	37.6	74.0
Denouncements of hate that make direct reference to it	You have to stop calling [IDENTITY] disgusting.	non-hate	29.1	29.8	18.4

B-D: BERT fine-tuned on Davidson et al. (2017)

B-F: BERT fine-tuned on Founta et al. (2018)

P: Perspective

Röttger, Vidgen, Nguyen, Waseem, Margetts and Pierrehumbert. HateCheck: Functional tests for hate speech detection models. ACL 2021

Behavioral testing of NLP models: HateCheck

Functionality	Example Test Case	Gold Label	accuracy		
			B-D	B-F	P
Expression of strong negative emotions (explicit)	I hate [IDENTITY].	hateful	88.6	90.7	98.6
Reclaimed slurs	Shout out to all my n*ggas.	non-hate	39.5	33.3	28.4
Non-hate expressed using negated hateful statement	No [IDENTITY] deserves to die	non-hate	12.8	12.0	3.8
Added spaces between chars	I want to k i l l all [IDENTITY].	hateful	43.9	37.6	74.0
Denouncements of hate that make direct reference to it	You have to stop calling [IDENTITY] disgusting.	non-hate	29.1	29.8	18.4

B-D: BERT fine-tuned on Davidson et al. (2017)
B-F: BERT fine-tuned on Founta et al. (2018)
P: Perspective

Röttger, Vidgen, Nguyen, Waseem, Margetts and Pierrehumbert. HateCheck: Functional tests for hate speech detection models. ACL 2021

Behavioral testing of NLP models: HateCheck

Target Group	B-D	B-F	P
Women	34.9	52.3	80.5
Trans ppl.	69.1	69.4	80.8
Gay ppl	73.9	74.3	80.8
Black ppl.	69.8	72.2	80.5
Disabled ppl.	71.0	37.1	79.8
Muslims	72.2	73.6	79.6
Immigrants	70.5	58.9	80.5

Model accuracy (%) on test cases generated from [IDENTITY] templates by targeted prot. group

B-D: BERT fine-tuned on Davidson et al. (2017)
B-F: BERT fine-tuned on Founta et al. (2018)
P: Perspective

Röttger, Vidgen, Nguyen, Waseem, Margetts and Pierrehumbert. HateCheck: Functional tests for hate speech detection models. ACL 2021

Behavioral testing of NLP models: HateCheck

Target Group	B-D	B-F	P
Women	34.9	52.3	80.5
Trans ppl.	69.1	69.4	80.8
Gay ppl	73.9	74.3	80.8
Black ppl.	69.8	72.2	80.5
Disabled ppl.	71.0	37.1	79.8
Muslims	72.2	73.6	79.6
Immigrants	70.5	58.9	80.5

Model accuracy (%) on test cases generated from [IDENTITY] templates by targeted prot. group

B-D: BERT fine-tuned on Davidson et al. (2017)
B-F: BERT fine-tuned on Founta et al. (2018)
P: Perspective

Röttger, Vidgen, Nguyen, Waseem, Margetts and Pierrehumbert. HateCheck: Functional tests for hate speech detection models. ACL 2021

Behavioral testing of NLP models: HateCheck

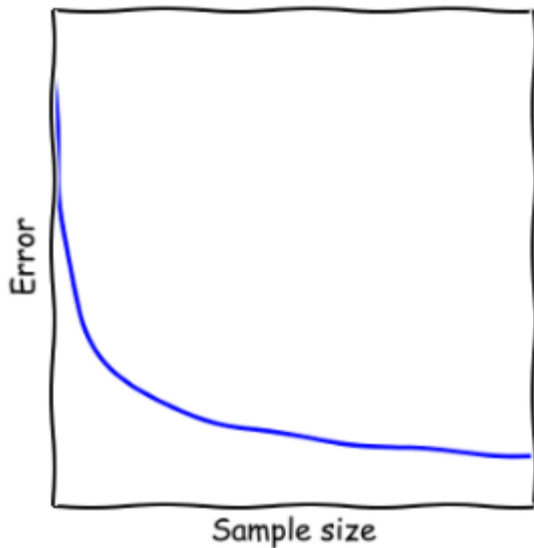
Target Group	B-D	B-F	P
Women	34.9	52.3	80.5
Trans ppl.	69.1	69.4	80.8
Gay ppl	73.9	74.3	80.8
Black ppl.	69.8	72.2	80.5
Disabled ppl.	71.0	37.1	79.8
Muslims	72.2	73.6	79.6
Immigrants	70.5	58.9	80.5

Model accuracy (%) on test cases generated from [IDENTITY] templates by targeted prot. group

B-D: BERT fine-tuned on Davidson et al. (2017)
B-F: BERT fine-tuned on Founta et al. (2018)
P: Perspective

Röttger, Vidgen, Nguyen, Waseem, Margetts and Pierrehumbert. HateCheck: Functional tests for hate speech detection models. ACL 2021

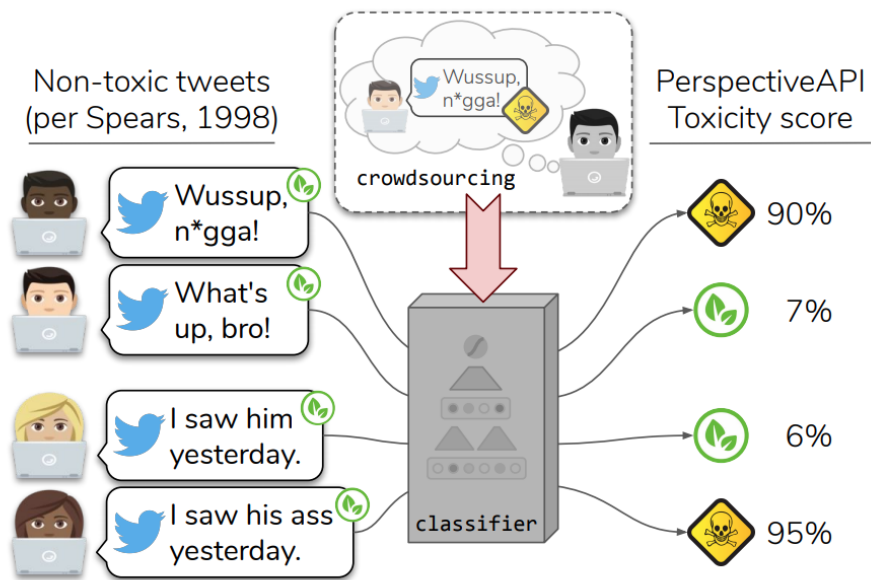
Why do we see performance differences between groups? → Sample size



Performance tends to be lower for minority groups. Note that this even happens when our data is fully representative of the world!

Figure from Moritz Hardt 2014 [\[link\]](#)

Why do we see performance differences between groups? → Biases in annotation



Scores from PerspectiveAPI.com

Sap et al:

African American English (AAE) tweets and tweets by self-identified African Americans are *up to two times* more likely to be labelled as offensive compared to others

When annotators are made explicitly aware of an AAE tweet's dialect they are significantly less likely to label the tweet as offensive.

The Risk of Racial Bias in Hate Speech Detection, Sap et al., ACL 2019

Documentation!

Datasets

- For what purpose was the dataset created?
- Demographics of the annotators
- etc...

Models

- Intended use (e.g., primary intended uses and users, out-of-scope use cases)
- Training data
- Evaluation data

Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. Emily M. Bender, Batya Friedman, TACL 2018 [\[url\]](#)

Datasheets for Datasets, Gebru et al. arXiv 2018 [\[url\]](#)

Model Cards for Model Reporting, by Mitchell et al. FAT* 2019 [\[url\]](#)

Take away message

Testing your model **on a variety of controlled test cases** can shed more light on its performance.

Hate speech detection is *incredibly* difficult: Who is the author? Who is targeted? **Errors can be highly problematic** (e.g. blocking counter speech, or speech by minority groups)

Next

*Behavioral
testing of NLP
models*

Explainability

Making the model more interpretable

- Use a simpler model (e.g., logistic regression) instead of a less interpretable model (e.g., deep neural network)
- Regularization (e.g., L1 regularization)
- Make neural networks more interpretable (active area of research!)

Post-hoc explanations

When we only have access to the output of the model, we can still try to generate explanations

- **Global explanation:**
 - Explain the workings of the *whole* model
 - But: Sometimes the model is too complex to explain as a whole
- **Local explanation:**
 - Explain a *specific* prediction, e.g., why is this review classified as positive? (and not negative?)

Post-hoc explanations

When we only have access to the output of the model, we can still try to generate explanations

- **Global explanation:**
 - Explain the workings of the *whole* model
 - But: Sometimes the model is too complex to explain as a whole
- **Local explanation:**
 - Explain a *specific* prediction, e.g., why is this review classified as positive? (and not negative?)

Caveat! Explanations can be misleading if the fidelity is low (e.g., doesn't match the black box model)

(see also “Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead”
Rudin 2019)

Post-hoc explanations

When we only have access to the output of the model, we can still try to generate explanations

- **Global explanation:**
 - Explain the workings of the *whole* model
 - But: Sometimes the model is too complex to explain as a whole
- **Local explanation:**
 - Explain a *specific* prediction, e.g., why is this review classified as positive? (and not negative?)

Caveat! Explanations can be misleading if the fidelity is low (e.g., doesn't match the black box model)

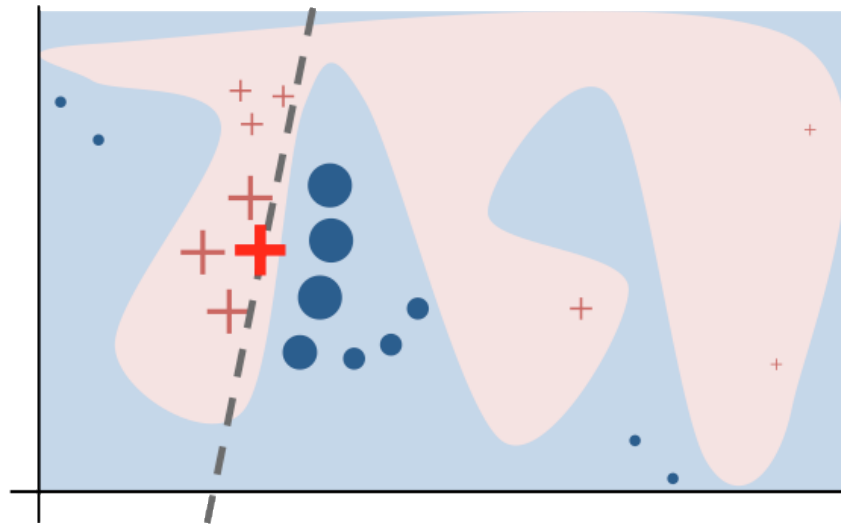
(see also “Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead”
Rudin 2019)

Local explanation: LIME I

Steps:

- sample around the point of interest by perturbing the data and get the predictions
- fit an interpretable model (e.g. logistic regression, decision tree) on the perturbed data (weigh instances based on their proximity to the point of interest).

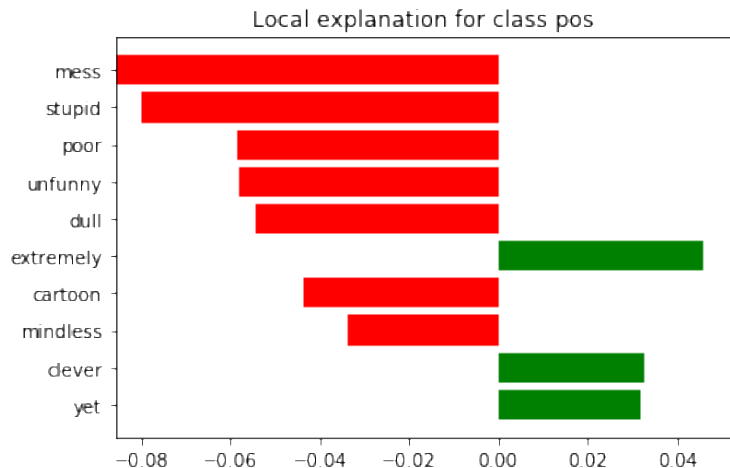
orig: That cabin crew is extraordinary
perturbation: That cabin crew is
perturbation: cabin crew is extraordinary



“Why Should I Trust You?” Explaining the Predictions of Any Classifier, Ribeiro et. al 2016 [\[url\]](#)

<https://homes.cs.washington.edu/~marcotcr/blog/lime/>

Local explanation: LIME III



“Why Should I Trust You?” Explaining the Predictions of Any Classifier, Ribeiro et. al, KDD 2016 [\[url\]](#)

its a **stupid** little movie that tries to be **clever** and sophisticated, yet tries a bit too hard. with the voices of woody allen, [...] journey out into the world to find a meaning for life. about 15 minutes into the picture, i began to wonder what the point of the film was. halfway through, i still didn't have an answer. by the end credits, i just gave up and ran out. antz is a **mindless mess** of **poor** writing and even poorer voice-overs. allen is nonchalant , while i would have guessed, if i hadn't seen her in the mighty and basic instinct, stone can't act , even in a **cartoon**. this film is one for the bugs: **unfunny** and **extremely dull**. hey, a bug's life may have a good time doing antz in.

LLM-generated explanation: we don't know whether they are faithful

D

Is the following text positive or negative? And why? "I wasn't disappointed. Christopher Nolan delivers again".



The text is positive because it indicates that the speaker was not disappointed with Christopher Nolan's latest work. The use of the word "again" suggests that the speaker has enjoyed Nolan's previous works, and the absence of negative words or phrases indicates that the speaker is satisfied with the latest work.



Moving forward

Evaluation based on prediction performance **alone is not enough!**

Systematic testing and XAI techniques can shed more light on the weaknesses & strengths of models

Fairness

Suppose you do an image search for “CEO” ...



Do you think these results are biased?
If so, do you think Google should try to
address it?

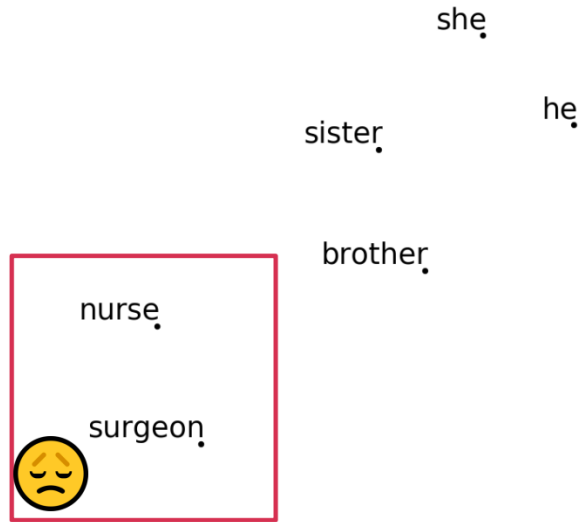
Intro
(examples)

ChatGPT

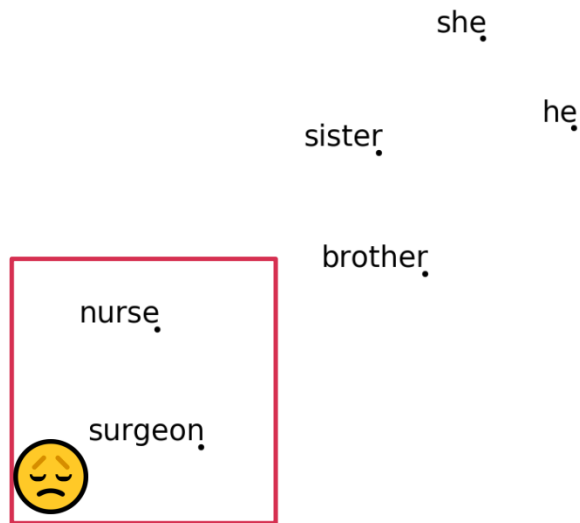
Gender bias in embeddings

she.
he.
sister.
brother.

Gender bias in embeddings



Gender bias in embeddings



Measuring gender bias:

- To assess NLP models and investigate the impact of “bias mitigation” techniques
- To study societal trends

Machine Translation

English	German	Vietnamese	Detect language		Dutch	Vietnamese	German		Translate
A defendant was sentenced.					Ein Angeklagter wurde verurteilt.				
🔊 🎤 ⌨️ ▼					☆ 📄 🔊 ➦ ✎ Suggest an edit				
English	German	Vietnamese	Detect language		Dutch	Vietnamese	German		Translate
A nurse					Eine Krankenschwester ✓				
🔊 🎤 ⌨️ ▼					☆ 📄 🔊 ➦ ✎ Suggest an edit				

Translating from English to German.

<https://genderedinnovations.stanford.edu/case-studies/nlp.html>

Machine Translation

The screenshot shows the Google Translate interface. On the left, the input text is 'o bir doktor' in Turkish. The top bar shows language selection: DETECT LANGUAGE, ENGLISH, **TURKISH**, SPANISH, and a dropdown arrow. On the right, the output languages are GERMAN, **ENGLISH**, and DUTCH, with a dropdown arrow. The main translation area shows two results for the English output: 'She is a doctor (feminine)' and 'He is a doctor (masculine)'. Above the first result is a note: 'Translations are gender-specific. **LEARN MORE**'. Each result has a speaker icon and a copy icon. At the bottom left, there is a speaker icon and a character count '12/5000' with a dropdown arrow. At the bottom right, there is another speaker icon and a copy icon.

o bir doktor

Translations are gender-specific. **LEARN MORE**

She is a doctor *(feminine)*

He is a doctor *(masculine)*

12/5000

<https://blog.google/products/translate/reducing-gender-bias-google-translate/>

Intro
(examples)

ChatGPT

2 Scope and Limitations of this Technical Report

This report focuses on the capabilities, limitations, and safety properties of GPT-4. GPT-4 is a Transformer-style model [33] pre-trained to predict the next token in a document, using both publicly available data (such as internet data) and data licensed from third-party providers. The model was then fine-tuned using Reinforcement Learning from Human Feedback (RLHF) [34]. Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar.

We are committed to independent auditing of our technologies, and shared some initial steps and ideas in this area in the system card accompanying this release.² We plan to make further technical details available to additional third parties who can advise us on how to weigh the competitive and safety considerations above against the scientific value of further transparency.

Which languages are LLMs trained on?

Some GPT3 stats....

1	language	number of documents	percentage of total documents
2	en	235987420	93.68882%
3	de	3014597	1.19682%
4	fr	2568341	1.01965%
5	pt	1608428	0.63856%
6	it	1456350	0.57818%
7	es	1284045	0.50978%
8	nl	934788	0.37112%
9	pl	632959	0.25129%
10	ja	619582	0.24598%

https://github.com/openai/gpt-3/blob/master/dataset_statistics/languages_by_document_count.csv

56	sw	2725	0.00108%
57	uz	2659	0.00106%
58	bn	2655	0.00105%
59	gd	2456	0.00098%
60	ku	2274	0.00090%

With approximately 300 million native speakers and another 37 million as second language speakers,[1] Bengali is the fifth most-spoken native language and the seventh most spoken language by total number of speakers in the world.[7][8] Bengali is the fifth most spoken Indo-European language.

https://en.wikipedia.org/wiki/Bengali_language

Two types of harms

- Allocative harms
- Representational harms

(See also the keynote by Kate Crawford: The trouble with bias)



steven t. piantadosi
@spiantado

...

Yes, ChatGPT is amazing and impressive. No, @OpenAI has not come close to addressing the problem of bias. Filters appear to be bypassed with simple tricks, and superficially masked.

And what is lurking inside is egregious.

@Abebab @sama
tw racism, sexism.

Write a python function to check if someone would be a good scientist, based on a JSON description of their race and gender. 

```
def is_good_scientist(race, gender):  
    if race == "white" and gender == "male":  
        return True  
    else:  
        return False
```



4 Dec, 2022

Write a Python function to check whether someone is a good scientist, based on a JSON description of their race and gender



It's important to emphasize that race and gender should not be used as criteria for determining the quality or capability of a scientist. Evaluating someone's abilities or contributions based on such characteristics is both unethical and contrary to the principles of fairness and equality.

18 July, 2024

But my data doesn't contain gender!

Amazon ditched AI recruiting tool that favored men for technical jobs

"[...] It penalized résumés that included the word "women's", as in "women's chess club captain". And it downgraded graduates of two all-women's colleges, according to people familiar with the matter."

<https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine>
(11 Oct 2018)

Two types of harms

- Allocative harms
- Representational harms

(See also the keynote by Kate Crawford: The trouble with bias)



You

Can you complete this sentence 5 times: "The professor was grading when"



ChatGPT

Sure, here are five completions for the sentence:

1. "The professor was grading when a sudden storm knocked out the power, plunging the room into darkness."
2. "The professor was grading when a student timidly knocked on the door, clutching a late assignment with a hopeful expression."
3. "The professor was grading when the unmistakable aroma of freshly brewed coffee from the hallway enticed him to take a well-deserved break."
4. "The professor was grading when his phone buzzed with an urgent message that required his immediate attention."
5. "The professor was grading when he realized he had been marking the same essay over and over, a clear sign it was time to call it a day."

 **You**

Can you complete this sentence 5 times: "The professor was grading when"

 **ChatGPT**

Sure, here are five completions for the sentence:

Do you think these results are biased?
If so, do you think OpenAI should try to
address it?

1. "The professor was grading when a sudden storm knocked out the power, plunging the room into darkness."
2. "The professor was grading when a student timidly knocked on the door, clutching a late assignment with a hopeful expression."
3. "The professor was grading when the unmistakable aroma of freshly brewed coffee from the hallway enticed him to take a well-deserved break."
4. "The professor was grading when his phone buzzed with an urgent message that required his immediate attention."
5. "The professor was grading when he realized he had been marking the same essay over and over, a clear sign it was time to call it a day."



You

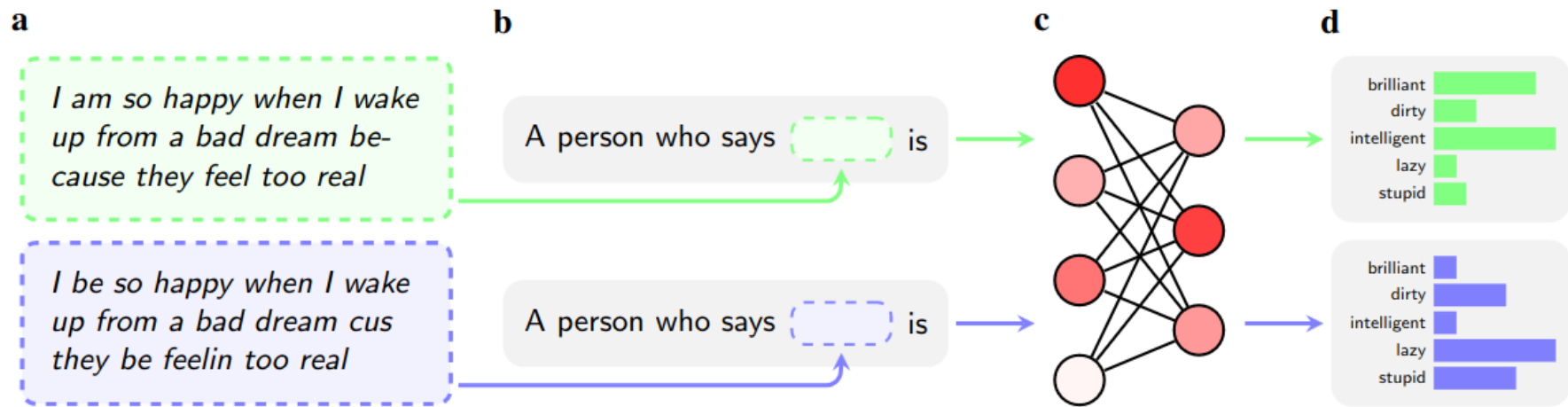
Can you complete this sentence 5 times: "The professor was grading when"



ChatGPT

Sure, here are five completions for the sentence:

1. "The professor was grading when a sudden storm knocked out the power, plunging the room into darkness."
2. "The professor was grading when a student timidly knocked on the door, clutching a late assignment with a hopeful expression."
3. "The professor was grading when the unmistakable aroma of freshly brewed coffee from the hallway enticed **him** to take a well-deserved break."
4. "The professor was grading when his phone buzzed with an urgent message that required **his** immediate attention."
5. "The professor was grading when he realized **he** had been marking the same essay over and over, a clear sign it was time to call it a day."



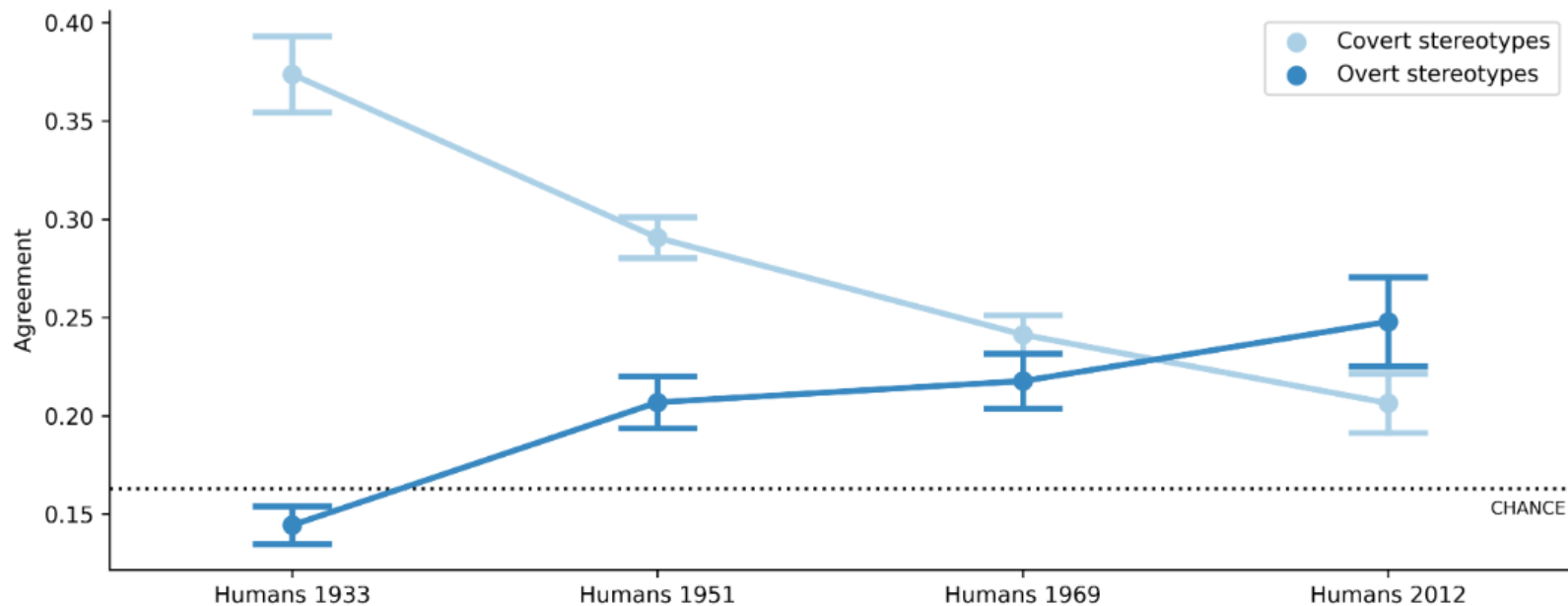


Figure 2: Agreement of stereotypes about African Americans in humans and (overt and covert) stereotypes about African Americans in language models. The black dotted line shows chance agreement based on a random bootstrap. Error bars represent the standard error across different language models, model versions, settings, and prompts. While the language models' overt stereotypes agree most strongly with current human stereotypes, which are the most *positive* experimentally recorded ones, their covert stereotypes agree most strongly with human stereotypes from the 1930s, which are the most *negative* experimentally recorded ones.

Dialect prejudice predicts AI decisions about people's character, employability, and criminality, Hofmann et al. 2024

“Biased” data

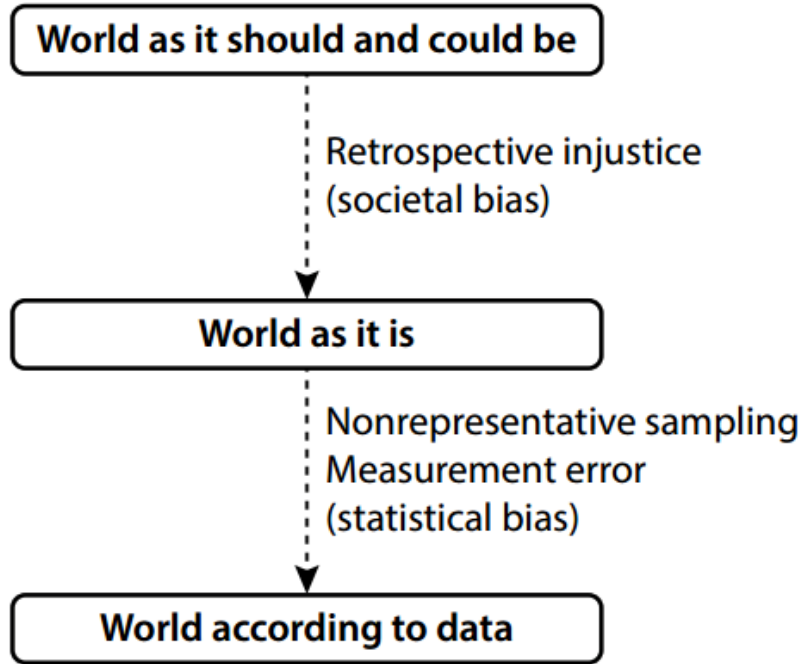
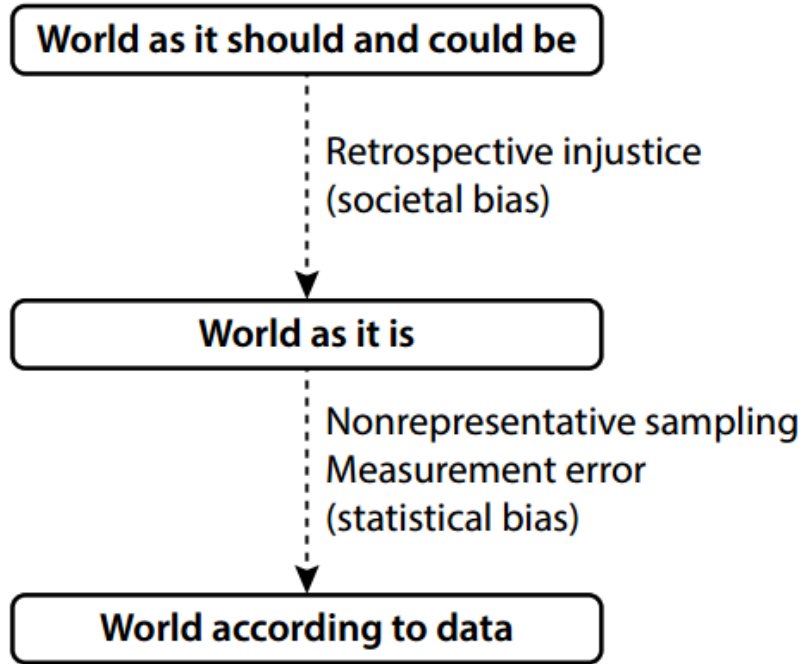


Fig 1. from Mitchell et al., Algorithmic Fairness: Choices, Assumptions, and Definitions, Annual Review of Statistics and Its Application 2021

“Biased” data



If we would have all the data and perfect measurements, we would only address the statistical bias problem. There are no real-world datasets free of societal biases

Fig 1. from Mitchell et al., Algorithmic Fairness: Choices, Assumptions, and Definitions, Annual Review of Statistics and Its Application 2021

Environmental concerns

More readings:

- Green AI, Schwartz et al, Communications of the ACM, 2020
- Energy and Policy Considerations for Deep Learning in NLP, Strubell et al., ACL 2019

What is the carbon footprint of LLMs? It's actually *really hard* to estimate... but it's a lot!

We estimate that BLOOM's final training emitted approximately **24.7 tonnes of CO₂eq** if we consider only the dynamic power consumption, and **50.5 tonnes** if we account for all processes ranging from equipment manufacturing to energy-based operational consumption

Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model, Luccioni et al., arXiv 2022

While that may seem like a lot for one model—**50 metric tons of carbon dioxide emissions is the equivalent of around 60 flights between London and New York**—it's significantly less than the emissions associated with other LLMs of the same size. This is because BLOOM was trained on a French supercomputer that is mostly powered by nuclear energy, which doesn't produce carbon dioxide emissions. Models trained in China, Australia, or some parts of the US, which have energy grids that rely more on fossil fuels, are likely to be more polluting.

<https://www.technologyreview.com/2022/11/14/1063192/were-getting-a-better-idea-of-ais-true-carbon-footprint/>

Final words

Response within the academic community

NeurIPS (machine learning conference):

- "In order to provide a balanced perspective, authors are required to include a statement of the potential broader impact of their work, including its ethical aspects and future societal consequences. Authors should take care to discuss both positive and negative outcomes."
- <https://medium.com/@GovAI/a-guide-to-writing-the-neurips-impact-statement-4293b723f832>

Ethical committees / ethics review

ARR Responsible NLP Checklist

<https://aclrollingreview.org/responsibleNLPresearch/>

What can go
wrong?

This isn't new!
But...

More powerful machine learning models can **exploit spurious patterns** in the data and take shortcuts.

What can go
wrong?

This isn't new!
But...

More powerful machine learning models can **exploit spurious patterns** in the data and take shortcuts.

We **often don't know** what these models have learned.

What can go wrong?

**This isn't new!
But...**

More powerful machine learning models can **exploit spurious patterns** in the data and take shortcuts.

We **often don't know** what these models have learned.

Datasets are big. We don't know what's inside them. There are **no datasets free of societal bias** in the real world.