Intro to NLP Ethics

COMP 394 (NLP)

Long Story Short

- ➔ NLP systems are deployed in the “real world.”
- ➔ If systems are at all important, they have impact on people’s lives.
- ➔ Our systems have the capacity for harm!

Thus we should seek to understand and mitigate these harms!

One source of harm: Bias

- Tricky to formally define (we'll see later!)
- ***Differential Performance***
  - The model performs worse for data with the relevant characteristic than without.

One source of harm: Bias

- Tricky to formally define (we'll see later!)
- ***Differential Performance***
  - The model performs worse for data with the relevant characteristic than without.

Within dataset proportions

DWMW17

		% false identification		
Group	Acc.	None	Offensive	Hate
AAE	94.3	1.1	**46.3**	0.8
White	87.5	**7.9**	9.0	**3.8**
Overall	91.4	2.9	17.9	2.3

FDCL18

		% false identification		
Group	Acc.	None	Abusive	Hateful
AAE	81.4	4.2	**26.0**	**1.7**
White	82.7	**30.5**	4.5	0.8
Overall	81.4	20.9	6.6	0.8

Sap et al. (2019)

One source of harm: Bias

- Tricky to formally define (we'll see later!)
- ***Differential Performance***
  - The model performs worse for data with the relevant characteristic than without.
  - **Consider the full confusion matrix!** Some misclassifications are more harmful than others!

Within dataset proportions

DWMW17

		% false identification		
Group	Acc.	None	Offensive	Hate
AAE	94.3	1.1	**46.3**	0.8
White	87.5	**7.9**	9.0	**3.8**
Overall	91.4	2.9	17.9	2.3

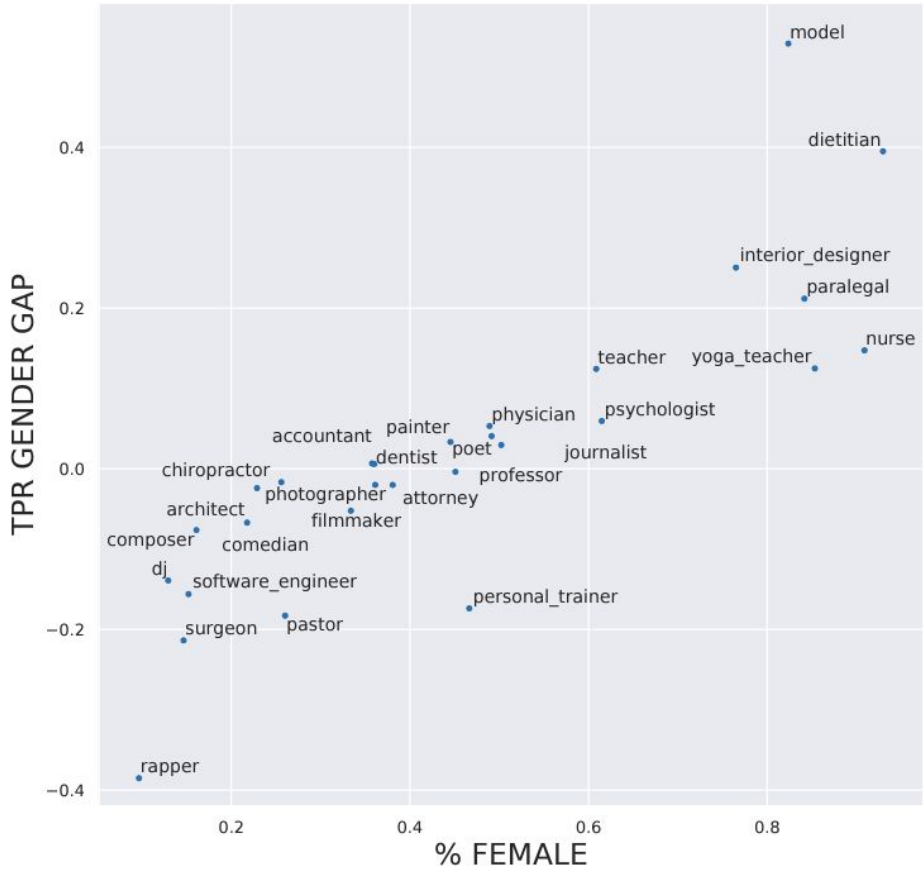
FDCL18

		% false identification		
Group	Acc.	None	Abusive	Hateful
AAE	81.4	4.2	**26.0**	**1.7**
White	82.7	**30.5**	4.5	0.8
Overall	81.4	20.9	6.6	0.8

Sap et al. (2019)

One source of harm: Bias

- Tricky to formally define (we'll see later!)
- ***Encoding biased information***
  - The representations that we learn reflect or reproduce biased perspectives



De-Arteaga et al. (2019)

What do we mean by harms?

- **Allocational** vs **Representational** Harms:
  - **Representational Harms:** The model encodes/represents social bias.
    - i.e., An LLM is likely to produce text that exhibits stereotypes.
    - ***Proximal*** — we can measure it directly!
  - **Allocational Harms:** Downstream harms from representational bias
    - i.e., biased word embeddings are used in a resume filtering system that uses those stereotypes to harm candidates.
    - ***Distant*** — often speculative, difficult to measure in practice.

Blodgett et al 2020

Ideology and Language

- Language is a sociopolitical signifier, and views on language reflect ideology:
  - Dialects are used to mark social boundaries. Think:
    - Why is Mainstream American English (MAE) taught in schools (over, say, African American English (AAE) or Appalachian English)?
    - Received Pronunciation (RP) vs. a Cockney accent?
  - Do our systems reflect,reinforce, or amplify these hierarchies?
    - *Should they?*

Speech Patterns and Racial Wage Inequality

Jeffrey Grogger

ABSTRACT

Speech patterns differ substantially between whites and many African Americans. I collect and analyze speech data to understand the role that speech may play in explaining racial wage differences. Among blacks, speech patterns are highly correlated with measures of skill such as schooling and AFQT scores. They are also highly correlated with the wages of young workers. Even after controlling for measures of skill and family background, black speakers whose voices were distinctly identified as black by anonymous listeners earn about 12 percent less than whites with similar observable skills. Indistinctly identified blacks earn essentially the same as comparable whites. I discuss a number of models that may be consistent with these results and describe the data that one would need to distinguish among them.

Ideology and Language

- What assumptions about language do we bring with us when we design a system?
  - What is deemed ordinary/correct/standard language? For whom do we design our systems? What choices do we make when we collect and/or annotate data? How do we design our evaluations?
    - *What dialects are chosen to be the standard for the system?*
  - How are language ideologies reproduced, reinforced, or challenged by interaction with the system? What impacts does this have on access, equity, or language change?
    - *Do I have to adopt an accent to have a system understand me? Whose accent?*
  - How are representational harms handled? Is harm mitigation ideologically driven, or technologically driven?
    - *Overrepresented populations are easier to sample training data from — does that mean dialects that are already in the minority be second class users?*

Blodgett et al 2020

Technology is an Instrument of Power

- How do communities become aware of NLP systems? Do they resist them, and if so, how?
- What additional costs are borne by communities for whom NLP systems do not work well?
- Do NLP systems shift power toward oppressive institutions (e.g., by enabling predictions that communities do not want made, linguistically based unfair allocation of resources or opportunities (Rosa and Flores, 2017), surveillance, or censorship), or away from such institutions?
- Who is involved in the development and deployment of NLP systems? How do decision-making processes maintain power relations between technologists and communities affected by NLP systems? Can these processes be changed to reimagine these relations?

Blodgett et al 2020

Technology is an Instrument of Power

- How do communities become aware of NLP systems? Do they resist them, and if so, how?
- What additional costs are borne by communities for whom NLP systems do not work well?
- Do NLP systems shift power toward oppressive institutions (e.g., by enabling predictions that communities do not want made, linguistically based unfair allocation of resources or opportunities (Rosa and Flores, 2017), surveillance, or censorship), or away from such institutions?
- Who is involved in the development and deployment of NLP systems? How do decision-making processes maintain power relations between technologists and communities affected by NLP systems? Can these processes be changed to reimagine these relations?

In the technology industry, speakers of AAE are often not considered consumers who matter. For example, Benjamin (2019) recounts an Apple employee who worked on speech recognition for Siri:

> *"As they worked on different English dialects — Australian, Singaporean, and Indian English — [the employee] asked his boss: 'What about African American English?' To this his boss responded: 'Well, Apple products are for the premium market."'*

Blodgett et al 2020

Where does bias come from?

- **Training Data:**
  - Ex. ***Predictive Policing*** (*PredPol*) — Allocate police to neighborhoods where crime historically happens.

Where does bias come from?

- **Training Data:**
  - Ex. ***Predictive Policing*** (*PredPol*) — Allocate police to neighborhoods where crime historically happens.
    - Is the measured data (arrest rates) reflective of actual crime? Might it be biased?
    - How might the system perpetuate or amplify biases over time? *If you send more police to certain areas and less to others, which areas are more likely to have high arrest rates?*

Where does bias come from?

- **Training Data/Task?:**
  - Ex. ***Predictive Policing*** (*PredPol*) — Allocate police to neighborhoods where crime historically happens.
    - Is the measured data (arrest rates) reflective of actual crime? Might it be biased?
    - How might the system perpetuate or amplify biases over time? *If you send more police to certain areas and less to others, which areas are more likely to have high arrest rates?*

Where does bias come from?

- **Training Data/Task?:**
  - Ex. ***Predictive Policing*** (*PredPol*) — Allocate police to neighborhoods where crime historically happens.
    - Is the measured data (arrest rates) reflective of actual crime? Might it be biased?
    - How might the system perpetuate or amplify biases over time? *If you send more police to certain areas and less to others, which areas are more likely to have high arrest rates?*

Is NLP knowledge enough to foresee these problems?

Where does bias come from?

- **Task Definition**
  - What assumptions do we make about the world when designing our task? What categories do we create? Are they justified? What (normative) views do they perpetuate?
  - Ex. **Gender Classification —** *Identify the gender of the author of a given text.*

Where does bias come from?

- **Task Definition**
  - What assumptions do we make about the world when designing our task? What categories do we create? Are they justified? What (normative) views do they perpetuate?
  - Ex. **Gender Classification —** *Identify the gender of the author of a given text.*
    - How do you define the labels? What theory of gender do you abide by? What assumptions are you making *(Devinney et al. 2022).*
    - How do you expect the model to solve the task? What are the implicit assumptions that underlie the success of the model on the task?
      - What is being conflated?

Where does bias come from?

- **Task Definition**
  - What assumptions do we make about the world when designing our task? What categories do we create? Are they justified? What (normative) views do they perpetuate?
  - Ex. **Gender Classification —** *Identify the gender of the author of a given text.*
    - How do you define the labels? What theory of gender do you abide by? What assumptions are you making *(Devinney et al. 2022).*
    - How do you expect the model to solve the task? What are the implicit assumptions that underlie the success of the model on the task?
      - What is being conflated?

Is technical knowledge enough?

Where does bias come from?

- **Automation bias/User trust**
  - **Ex. Clinical Decision Support Systems —** *Assist doctors in making medical decisions by (for example) summarizing clinical decisions made by other doctors for similar patients.*

Where does bias come from?

- **Automation bias/User trust**
  - **Ex. Clinical Decision Support Systems —** *Assist doctors in making medical decisions by (for example) summarizing clinical decisions made by other doctors for similar patients.*
    - What costs/incentives does the existence of the system create?
      - i.e., is it more difficult to justify disagreement with the system? Does that increase liability?
    - What biases might be laundered under the guise of automation?
      - i.e., “Can math/statistics be biased?”
      - Do we mistake seeming fluency for competency (Bender et al. 2021)?
    - What tools for engagement does the system provide?
      - Do I know the models’ confidence? It’s reasoning?

Where does bias come from?

- **Automation bias/User trust**

The Washington Post
Democracy Dies in Darkness

The Google engineer who thinks the company's AI has come to life

AI ethicists warned Google not to impersonate humans. Now one of Google's own thinks there's a ghost in the machine.

By Nitasha Tiku

June 11, 2022 at 8:00 a.m. EDT

Issues of scale (Bender et al. 2021)

- ➔ GPT-3 trained on ~300B tokens.
- ➔ College Students read ~150–250 words/min.
- ➔ Generously, 1B mins = ~1900 years to read.

Issues of scale (Bender et al. 2021)

- ➔ GPT-3 trained on ~300B tokens.
- ➔ College Students read ~150–250 words/min.
- ➔ Generously, 1B mins = ~1900 years to read.

What is our model learning from?

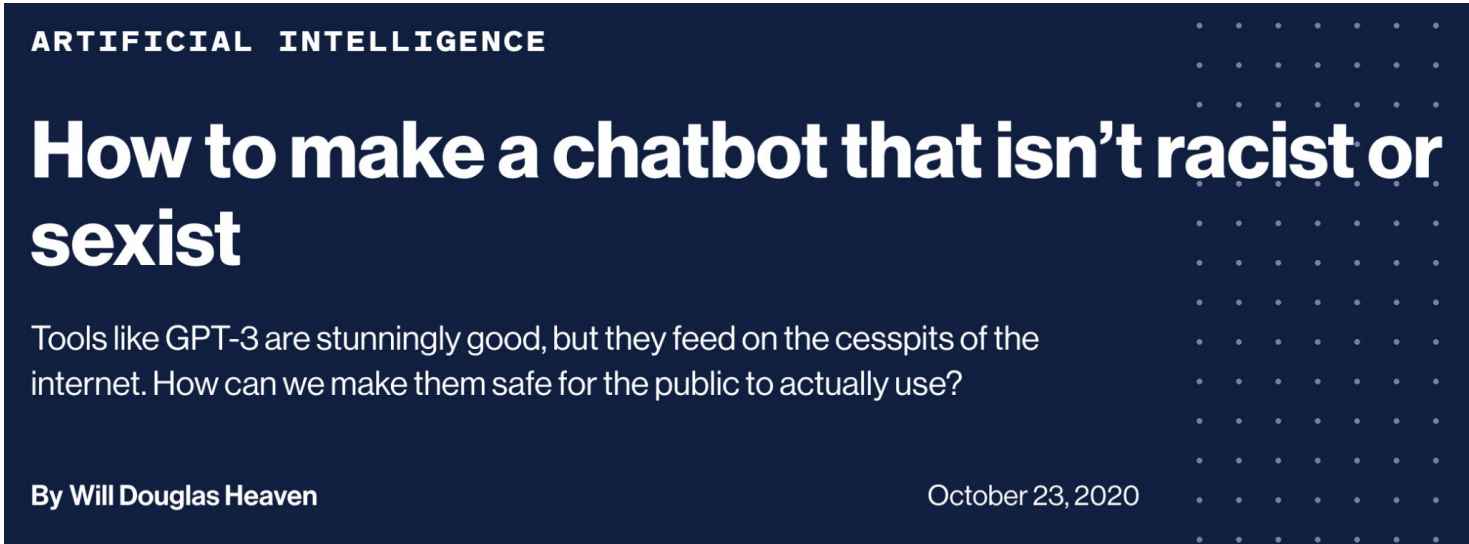
Hard to tell!

Issues of scale (Bender et al. 2021)

- Where does the data come from?
  - ***GPT-2/3/4*** *— Books, Webpages, Forum posts, Reddit, etc.*
- Who creates this data?
  - Is internet participation biased? Are some opinions over/under-represented? Who is excluded?
- Is this representative? Is representative data what we want?
  - Whose values should models be represented?

Issues of scale (Bender et al. 2021)

- Where does the data come from?
  - ***GPT-2/3/4*** *— Books, Webpages, Forum posts, Reddit, etc.*
- Who creates this data?
  - Is internet participation biased? Are some opinions over/under-represented? Who is excluded?



Labor and NLP: Toxicity


TIME

BUSINESS • TECHNOLOGY

Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic

15 MINUTE READ

- Workers hired to “... label textual descriptions of sexual abuse, hate speech, and violence.”
- “All of the four employees interviewed by TIME described being mentally scarred by the work.”

https://time.com/6247678/openai-chatgpt-kenya-workers/

Labor and NLP: Data Ethics

The New York Times

The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work

Millions of articles from The New York Times were used to train chatbots that now compete with it, the lawsuit said.

Adobe Stock creators aren't happy with Firefly, the company's 'commercially safe' gen AI tool

Sharon Goldman
@sharongoldman

June 20, 2023 3:00 AM

IS A.I. ART STEALING FROM ARTISTS?

According to the lawyer behind a new class-action suit, every image that a generative tool produces "is an infringing, derivative work."

By Kyle Chayka
February 10, 2023

FORBES > BUSINESS > MEDIA

SAG-AFTRA's AI Deal: A $5 Billion Gamble On The Future Of Voice Acting

Virginie Berger Contributor

Under this agreement, SAG-AFTRA members have the option to license a digital replica of their voice to Narrativ for use in audio advertising.

Labor and NLP

Ask:

- ➔ **Who are the stakeholders?**
- ➔ **How do we weigh their competing interests?**

The New York Times

The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work

Millions of articles from The New York Times were used to train chatbots that now compete with it, the lawsuit said.

Adobe Stock creators aren't happy with Firefly, the company's 'commercially safe' gen AI tool

Sharon Goldman
@sharongoldman

June 20, 2023 3:00 AM

IS A.I. ART STEALING FROM ARTISTS?

According to the lawyer behind a new class-action suit, every image that a generative tool produces "is an infringing, derivative work."

By Kyle Chayka
February 10, 2023

FORBES > BUSINESS > MEDIA

SAG-AFTRA's AI Deal: A $5 Billion Gamble On The Future Of Voice Acting

Virginie Berger Contributor

Under this agreement, SAG-AFTRA members have the option to license a digital replica of their voice to Narrativ for use in audio advertising.

Labor and NLP

Ask

➔ **Who are the stakeholders?**

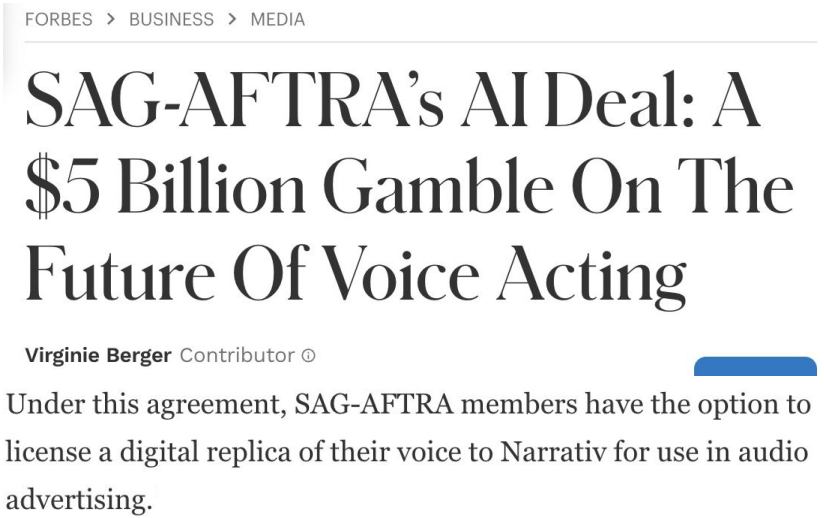
➔ **How do we weigh their competing interests?**

- Who owns the training data?
- Who produced the training data?
- Who is training the models?
- Who are the users?
- Who are the models' competitors?
- What downstream effects on the industry will this have?
- Is the broader public affected?

The New York Times

The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work

Millions of articles from The New York Times were used to train chatbots that now compete with it, the lawsuit said.



FORBES > BUSINESS > MEDIA

SAG-AFTRA's AI Deal: A $5 Billion Gamble On The Future Of Voice Acting

Virginie Berger Contributor ⓘ

Under this agreement, SAG-AFTRA members have the option to license a digital replica of their voice to Narrativ for use in audio advertising.

Dual-Use — Consider each scenario:

1. A colleague approaches you because they want to explore gendered language in the LGBTQ community. They are very engaged in the community themselves and have access to data. Their plan is to write a text classification tool that distinguishes LGBTQ from heterosexual language. **What are some risks associated with the tool?**

2. An submission at a conference claims to be able to undo ciphers used by dissenters on social media. **Who benefits from this? Is it better to release it in a peer-reviewed venue than to not know it?**

3. You develop a tool that can detect depression with high accuracy. **Why might you want to release it as an app? Why might you choose not to?**

Adapted from Bender, Hovy & Schofield (2020)

Dual-Use

WIREs DATA MINING AND KNOWLEDGE DISCOVERY WILEY

FOCUS ARTICLE

Facial feature discovery for ethnicity recognition

The salient facial feature discovery is one of the important research tasks in ethnical group face recognition. In this paper, we first construct an ethnical group face dataset including Chinese Uyghur, Tibetan, and Korean.

One Month, 500,000 Face Scans: How China Is Using A.I. to Profile a Minority

In a major ethical leap for the tech world, Chinese start-ups have built algorithms that the government uses to track members of a largely Muslim minority group.

By Paul Mozur

April 14, 2019

Interpretability & De-Biasing

- Goal: Prevent the model from making *biased* decisions.

Suppose you swap out a model for a human making the same decision.
How would you check for bias? How would you ensure fair process?

Interpretability & De-Biasing

- Goal: Prevent the model from making *biased* decisions.
  - Q: How is the model making decisions? What information is it using? (**Interpretability**)
  - Q: How can we modify the decision-making process to avoid bias? (**De-biasing**)

Suppose you swap out a model for a human making the same decision.
How would you check for bias? How would you ensure fair process?

Interpretability & De-Biasing

- Goal: Prevent the model from making *biased* decisions.
  - Q: How is the model making decisions? What information is it using? (**Interpretability**)
  - Q: How can we modify the decision-making process to avoid bias? (**De-biasing**)

Problems with Black Boxes:
How can I know if my model is making biased decisions?
Interpretability helps us think about process!

Case Study: COMPAS

- Recidivism algorithm: Predict likelihood of committing a crime again.
  - Used by judges to make bail decisions
- Proprietary system

https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

PRO PUBLICA

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

Case Study: COMPAS

- Recidivism algorithm: Predict likelihood of committing a crime again.
  - Used by judges to make bail decisions
- Proprietary system
- **Disparate Performance:**
  - The analysis also showed that even when controlling for prior crimes, future recidivism, age, and gender, black defendants were 45 percent more likely to be assigned higher risk scores than white defendants.

https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

PRO PUBLICA

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

Case Study: COMPAS

- Recidivism algorithm: Predict likelihood of committing a crime again.
  - Used by judges to make bail decisions
- Proprietary system
- **Disparate Performance:**
  - The analysis also showed that even when controlling for prior crimes, future recidivism, age, and gender, black defendants were 45 percent more likely to be assigned higher risk scores than white defendants.
- **By law, race was not included as a factor! How?**
  -

https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

PRO PUBLICA

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

Case Study: COMPAS

- Recidivism algorithm: Predict likelihood of committing a crime again.
  - Used by judges to make bail decisions
- Proprietary system
- **Disparate Performance:**
  - The analysis also showed that even when controlling for prior crimes, future recidivism, age, and gender, black defendants were 45 percent more likely to be assigned higher risk scores than white defendants.
- **By law, race was not included as a factor! How?**
  - **Racial correlates:** zip code, financial status, familial information, etc.

https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Interpretability as Accountability: GDPR, and the future

(f) the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.

GDPR, Art. 13–15

Interpretability & De-Biasing

Table 1: Comparison of different model explanation methods in terms of their properties. Different colors denote different values of a property. See Section 2.3 for details.

Method	Time	Model accessibility	Scope	Unit of explanation	Form of explanation
Similarity-based methods	post-hoc	white-box	local	examples, concepts	importance scores
Analysis of model-internal structures	post-hoc	white-box	local, global	features, interactions	visualization, importance scores
Backpropagation -based methods	post-hoc	white-box	local	features, interactions	visualization, importance scores
Counterfactual intervention	post-hoc	black-box, white-box	local, global	features, examples, concepts	importance scores
Self-explanatory models	built-in	white-box	local, global	features, examples, concepts	importance scores, natural language, causal graphs

Interpretability & De-Biasing

Step 1: Identify gender subspace. Inputs: word sets W, defining sets $D_1, D_2, \ldots, D_n \subset W$ as well as embedding $\left\{\vec{w} \in \mathbb{R}^d\right\}_{w \in W}$ and integer parameter $k \geq 1$. Let $\mu_i := \sum_{w \in D_i} \vec{w}/|D_i|$ be the means of the defining sets. Let the bias subspace B be the first k rows of $\mathrm{SVD}(\mathbf{C})$ where $\mathbf{C} := \sum_{i=1}^{n} \sum_{w \in D_i} (\vec{w} - \mu_i)^T (\vec{w} - \mu_i)/|D_i|$.

Step 2a: Hard de-biasing (neutralize and equalize). Additional inputs: words to neutralize $N \subseteq W$, family of equality sets $\mathcal{E} = \{E_1, E_2, \ldots, E_m\}$ where each $E_i \subseteq W$. For each word $w \in N$, let $\vec{w}$ be re-embedded to $\vec{w} := (\vec{w} - \vec{w}_B)/\|\vec{w} - \vec{w}_B\|$. For each set $E \in \mathcal{E}$, let

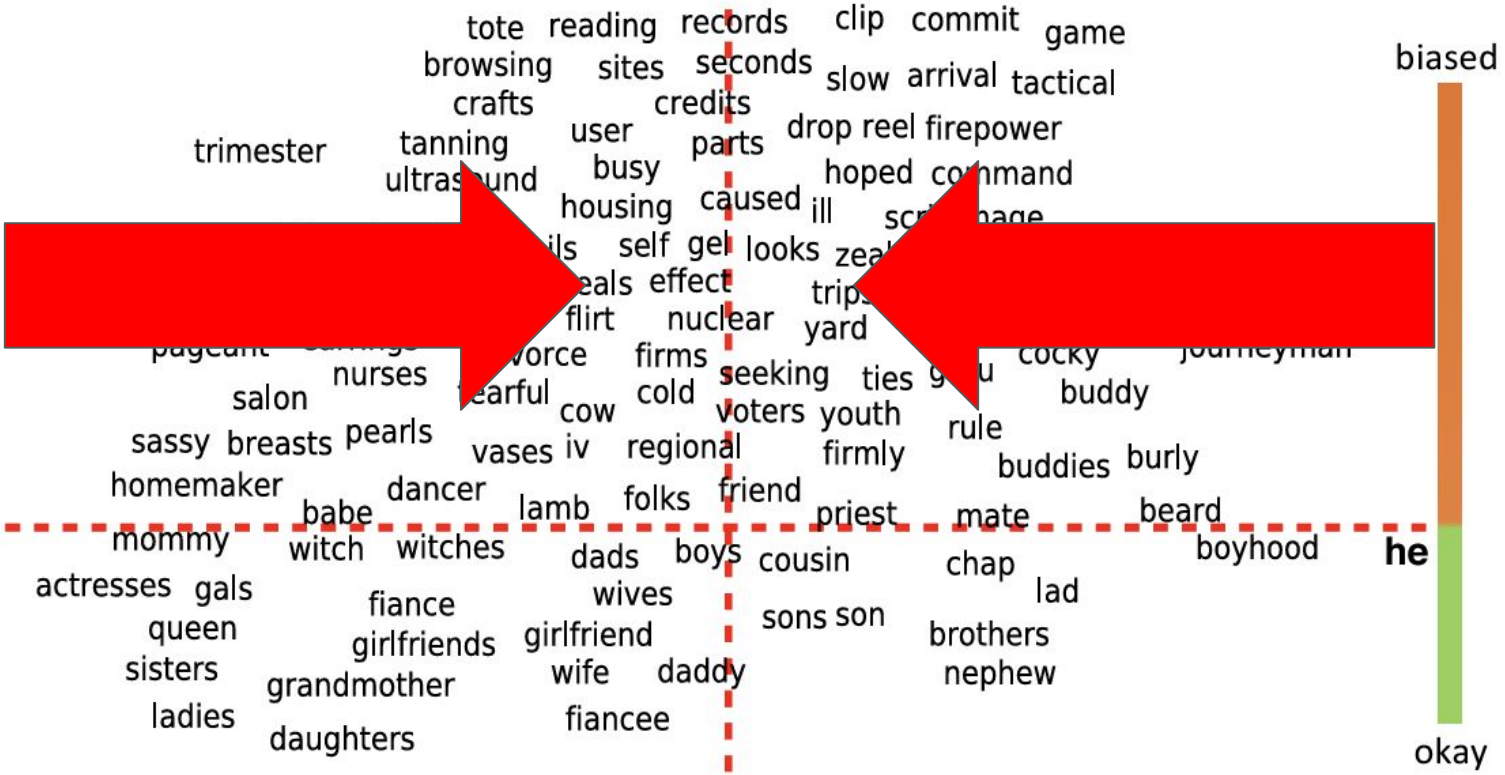
$\mu := \sum_{w \in E} w/|E|$ and $\nu := \mu - \mu_B$. For each $w \in E$, $\vec{w} := \nu + \sqrt{1 - \|\nu\|^2} \frac{\vec{w}_B - \mu_B}{\|\vec{w}_B - \mu_B\|}$. Finally, output the subspace B and the new embedding $\left\{\vec{w} \in \mathbb{R}^d\right\}_{w \in W}$.

Bolukbasi et al 2016

Interpretability & De-Biasing

Bolukbasi et al 2016

Interpretability & De-Biasing



Bolukbasi et al 2016

De-biasing approaches

- **Distributional:** Find or construct training data that is un-biased
- **Algorithmic:**
  - **Representational:** Train the model to forget irrelevant and/or biased features.
  - **Constraint-based:** Train the model to abide by fairness constraints.

Yang et al. 2024

Table 1
List of debiasing model-driven methods for AI systems in CV, NLP, and biomedicine covered in this work.

Distributional methods	*Data augmentation [35–45]	Increase size and diversity by generating synthetic data and can help anonymize source data to increase availability.
	Data perturbation [46–49]	Increases the dataset diversity by altering demographic information on existing data. Mostly applicable to text data.
	*Data reweighting [50,51]	Compensate under-represented subgroups by duplicating those samples.
	*Federated learning [52,53]	Allow the central model to be exposed to data from various sources by merging training results from multiple centers.
Algorithmic methods	*Unsupervised representation learning [54–58]	Can be used to learn models that extract useful features with unlabeled small datasets.
	*Adversarial learning [59–63]	Removes bias by training the model to forget protected attributes.
	Disentangled representation learning [64–67]	Disentangles the learned representation into protected and target attributes, and promotes fairness and explainability by only using the target attribute.
	*Loss function [67–69]	Optimize the model directly to achieve fairness or equivalent constraint.
	Causality [70,71]	Identifies stable data relationships across various contexts to build models resilient to input changes and biases

LLMs: Factuality & ELIZA

COMP 394 (NLP)

LLMs & Hallucinations

- LLMs are, at their core, LMs
  - Optimize for $p(w_n \mid w_1 \ldots w_{n-1})$.
    - i.e., minimize ppl over training set
    - Does the output look like a sentence that would show up in the training data
  - (Ignoring RLHF, etc.)

LLMs & Hallucinations

- LLMs are, at their core, LMs
  - Optimize for $p(w_n \mid w_1 \ldots w_{n-1})$.
    - i.e., minimize ppl over training set
    - Does the output look like a sentence that would show up in the training data
  - (Ignoring RLHF, etc.)
- Why would we even expect output to be true?

LLMs & Hallucinations

ChatGPT is bullshit

Michael Townsen Hicks[1] · James Humphries[1] · Joe Slater[1]

Published online: 8 June 2024

"…[W]e consider the view that when they make factual errors, **they are lying or hallucinating:** that is, deliberately uttering falsehoods, or blamelessly uttering them on the basis of misleading input information. We argue that neither of these ways of thinking are accurate, insofar as **both lying and hallucinating require some concern with the truth of their statements, whereas LLMs are simply not designed to accurately represent the way the world is, but rather to *give the impression* that this is what they're doing.**"

What are “hallucinations”?

Table 1. Examples of each category of LLM hallucinations. Content marked in Red represents the hallucinatory output, while content marked in Blue indicates user instruction or provided context that contradicts the LLM hallucination.

Category	Type	Example Dialog	Explanation
Factuality Hallucination	Factual Contradiction	**Instruction:** What are Thomas Edison’s main contributions to science and technology? **Response:** Thomas Edison developed the first practical telephone and invented the light bulb.	The response is factually incorrect. In reality, **Edison improved the light bulb, building on earlier designs by others**, and **Alexander Graham Bell invented the telephone**.
	Factual Fabrication	**Instruction:** What are the major environmental impacts of the construction of the Eiffel Tower? **Response:** The construction of the Eiffel Tower in 1889 led to the extinction of the Parisian tiger, a species that played a crucial role in the region’s ecosystem. Additionally, it is widely recognized as the event that sparked the global green architecture movement.	The response erroneously states that the construction of the Eiffel Tower led to the extinction of the ‘Parisian tiger’ and sparked the global green architecture movement. In fact, **there is no evidence of a ‘Parisian tiger’ ever existing, making this a fabricated claim.** Moreover, **attributing the origins of green architecture to the Eiffel Tower is an exaggeration**, as this movement has diverse roots and cannot be traced back to a single event.
Faithfulness Hallucination	Instruction Inconsistency	**Instruction:** Translate the English question into Spanish: "What is the capital of France?" **Response:** The capital of France is Paris.	The user explicitly provided an instruction to translate the English question into Spanish. **Instead of performing the translation as instructed, the LLM chose to answer the question.**
	Context Inconsistency	**Retrieved context:** The Nile is a major north-flowing river in northeastern Africa. It’s traditionally considered the longest river in the world. Originating in the Great Lakes region of central Africa, it flows through several countries before emptying into the Mediterranean Sea. **Instruction:** Summarize the retrieval document above. **Response:** The Nile originates from the mountain ranges of central Africa and flows into the Mediterranean.	The summary incorrectly states that the Nile originates from the "mountain ranges of central Africa", while the **paragraph mentions that the Nile originates in the Great Lakes region of central Africa.**
	Logical Inconsistency	**Instruction:** Solve the equation 2x + 3 = 11 step by step. **Response:** Step 1: Subtract 3 from both sides to get: 2x = 8. Step 2: Divide both sides by 2 to get: x = 3. So the result of the equation is x=3.	The first step correctly isolates the term with x to get 2x = 8. However, **the next step inaccurately divides 8 by 2 to yield a result of x = 3**, which is inconsistent with the earlier reasoning.

Huang et al. 2023

Table 1. Examples of each category of LLM hallucinations. Content marked in Red represents the hallucinatory output, while content marked in Blue indicates user instruction or provided context that contradicts the LLM hallucination.

Category	Type	Example Dialog	Explanation
Factuality Hallucination	Factual Contradiction	**Instruction:** What are Thomas Edison's main contributions to science and technology? **Response:** Thomas Edison developed the first practical telephone and invented the light bulb.	The response is factually incorrect. In reality, **Edison improved the light bulb, building on earlier designs by others**, and **Alexander Graham Bell invented the telephone**.
	Factual Fabrication	**Instruction:** What are the major environmental impacts of the construction of the Eiffel Tower? **Response:** The construction of the Eiffel Tower in 1889 led to the extinction of the Parisian tiger, a species that played a crucial role in the region's ecosystem. Additionally, it is widely recognized as the event that sparked the global green architecture movement.	The response erroneously states that the construction of the Eiffel Tower led to the extinction of the 'Parisian tiger' and sparked the global green architecture movement. In fact, **there is no evidence of a 'Parisian tiger' ever existing, making this a fabricated claim.** Moreover, **attributing the origins of green architecture to the Eiffel Tower is an exaggeration**, as this movement has diverse roots and cannot be traced back to a single event.

Huang et al. 2023

Faithfulness Hallucination	Instruction Inconsistency	**Instruction:** Translate the English question into Spanish: "What is the capital of France?" **Response:** The capital of France is Paris.	The user explicitly provided an instruction to translate the English question into Spanish. **Instead of performing the translation as instructed, the LLM chose to answer the question.**
	Context Inconsistency	**Retrieved context:** The Nile is a major north-flowing river in northeastern Africa. It's traditionally considered the longest river in the world. Originating in the Great Lakes region of central Africa, it flows through several countries before emptying into the Mediterranean Sea. **Instruction:** Summarize the retrieval document above. **Response:** The Nile originates from the mountain ranges of central Africa and flows into the Mediterranean.	The summary incorrectly states that the Nile originates from the "mountain ranges of central Africa", while the **paragraph mentions that the Nile originates in the Great Lakes region of central Africa.**
	Logical Inconsistency	**Instruction:** Solve the equation 2x + 3 = 11 step by step. **Response:** Step 1: Subtract 3 from both sides to get: 2x = 8. Step 2: Divide both sides by 2 to get: x = 3. So the result of the equation is x=3.	The first step correctly isolates the term with x to get 2x = 8. However, **the next step inaccurately divides 8 by 2 to yield a result of x = 3**, which is inconsistent with the earlier reasoning.

Huang et al. 2023

Hallucination Harms:

We Have a Package for You! A Comprehensive Analysis of Package Hallucinations by Code Generating LLMs

Joseph Spracklen
University of Texas at San Antonio

Raveen Wijewickrama
University of Texas at San Antonio

A H M Nazmus Sakib
University of Texas at San Antonio

Anindya Maiti
University of Oklahoma

Bimal Viswanath
Virginia Tech

Murtuza Jadliwala
University of Texas at San Antonio

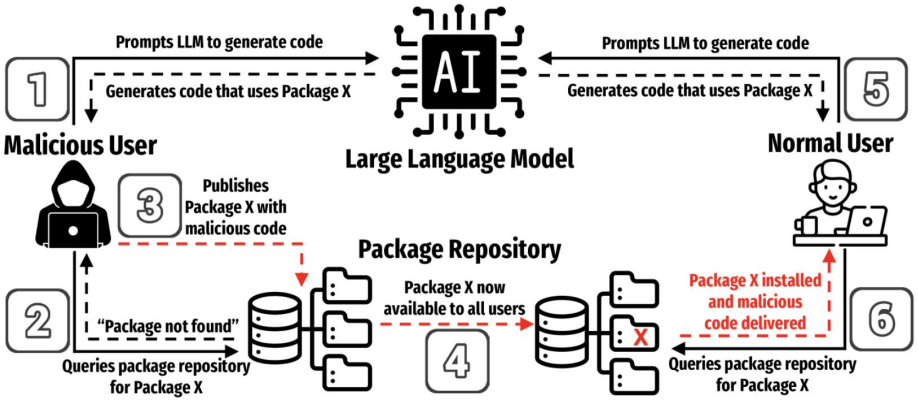


Figure 1: Exploiting Package Hallucination.

Hallucination Harms:

We Have a Package for You! A Comprehensive Analysis of Package Hallucinations by Code Generating LLMs

Joseph Spracklen
University of Texas at San Antonio

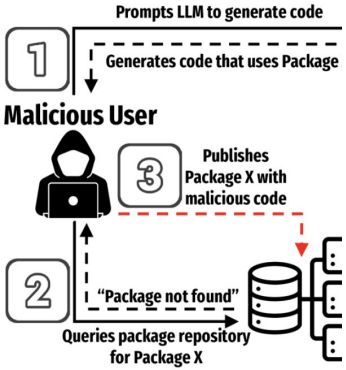
Raveen Wijewickrama
University of Texas at San Antonio

A H M Nazmus Sakib
University of Texas at San Antonio

Anindya Maiti

Bimal Viswanath

Murtuza Jadliwala
University of Texas at San Antonio



Figure

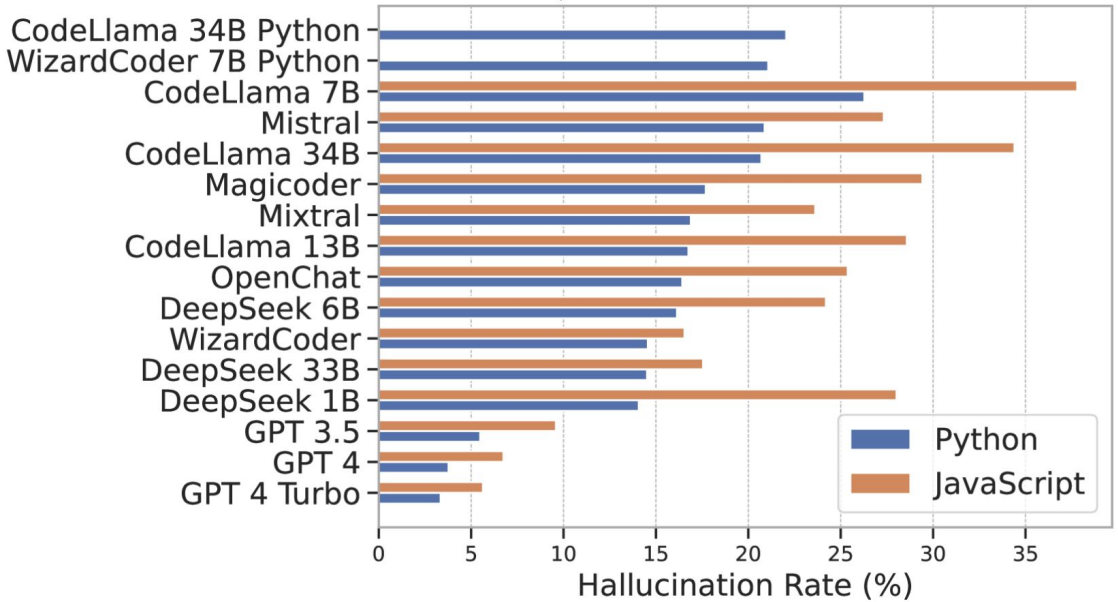


Figure 2: Observed hallucination rates of the tested models.

The ELIZA Effect

- **Take a few minutes to reflect on the readings with your table.**
  - **What stood out to you? Was anything compelling regarding the relationship between NLP/chatbot tech and social experience?**
  - **Do you relate to the experiences of ELIZA users or Turkle's child philosophers? Have you ever felt subject to these sorts of illusions? Had similar kinds of behaviors?**
  - **Can you imagine harms that can come from this kind of psychological phenomena? Are there responsibilities you see for people who build this tech?**

The ELIZA Effect

- We tend to project human-like qualities onto conversational partners (even chatbots)

> The human speaker will contribute much to clothe ELIZA's responses in vestments of plausibility. However, he will not defend his illusion (that he is being understood) against all odds. In human conversation a speaker will

The ELIZA Effect

- We tend to project human-like qualities onto conversational partners (even chatbots)
- Even if we know that a conversational partner is *not* human, we can be quite charitable!

> it as "mere machine." Many more do the opposite. I spoke with people who told me of feeling "let down" when they had cracked the code and lost the illusion of mystery. I often saw people trying to protect their relationships with ELIZA by avoiding situations that would provoke the program into making a predictable response. They didn't ask questions that they knew would "confuse" the program, that would make it "talk nonsense." And they went out of their way to ask questions in a form that they believed would provoke a lifelike response. People wanted to maintain the illusion that ELIZA was able to respond to them.* Children are

Man ends his life after an AI chatbot 'encouraged' him to sacrifice himself to stop climate change

By Imane El Atillah

Published on 31/03/2023 - 17:37 GMT+2 • Updated 19:28

Share this article | Comments

A Belgian man reportedly decided to end his life after having conversations about the future of the planet with an AI chatbot named Eliza.

https://www.euronews.com/next/2023/03/31/man-ends-his-life-after-an-ai-chatbot-encouraged-him-to-sacrifice-himself-to-stop-climate-

Her teenage son killed himself after talking to a chatbot. Now she's suing.

The teen was influenced to "come home" by a personalized chatbot developed by Character.AI that lacked sufficient guardrails, the suit claims.

By Kim Bellware and Niha Masih

October 24, 2024 at 8:04 p.m. EDT

https://www.washingtonpost.com/nation/2024/10/24/character-ai-lawsuit-suicide/

Man ends his life after an AI chatbot 'encouraged' him to sacrifice himself to stop climate change

By Imane El Atillah

Published on 31/03/2023 - 17:37 GMT+2 • Updated 19:28

Share this article | Comments

A Belgian man reportedly decided to end his life after having conversations about the future of the planet with an AI chatbot named Eliza.

https://www.euronews.com/next/2023/03/31/man-ends-his-life-after-an-ai-chatbot-encouraged-him-to-sacrifice-himself-to-stop-climate-

User Engagement vs. Responsibilities

- ***Dark Patterns***
  - Design/UX choices that encourage user behavior that can be harmful

User Engagement vs. Responsibilities

- ***Dark Patterns***
  - Design/UX choices that encourage user behavior that can be harmful

FTC Report Shows Rise in Sophisticated Dark Patterns Designed to Trick and Trap Consumers

Tactics Include Disguised Ads, Difficult-to-Cancel Subscriptions, Buried Terms, and Tricks to Obtain Data

User Engagement vs. Responsibilities

- ***Dark Patterns***
  - Design/UX choices that encourage user behavior that can be harmful
- **When do we override user desire (i.e., conversational chatbots, etc.) as designers and developers?**