

Bias in NLP Systems: Part I – How it Arises

Ethics in NLP: Including Society in Discourse and Design

February 21st, 2019

Ryan Georgi

OpenAI: Language Models, SciComm, and Ethics

- [Original Post](#) - [[code](#)] [[tech report](#)]
- Headlines:

OpenAI: Language Models, SciComm, and Ethics

- [Original Post](#) - [[code](#)] [[tech report](#)]

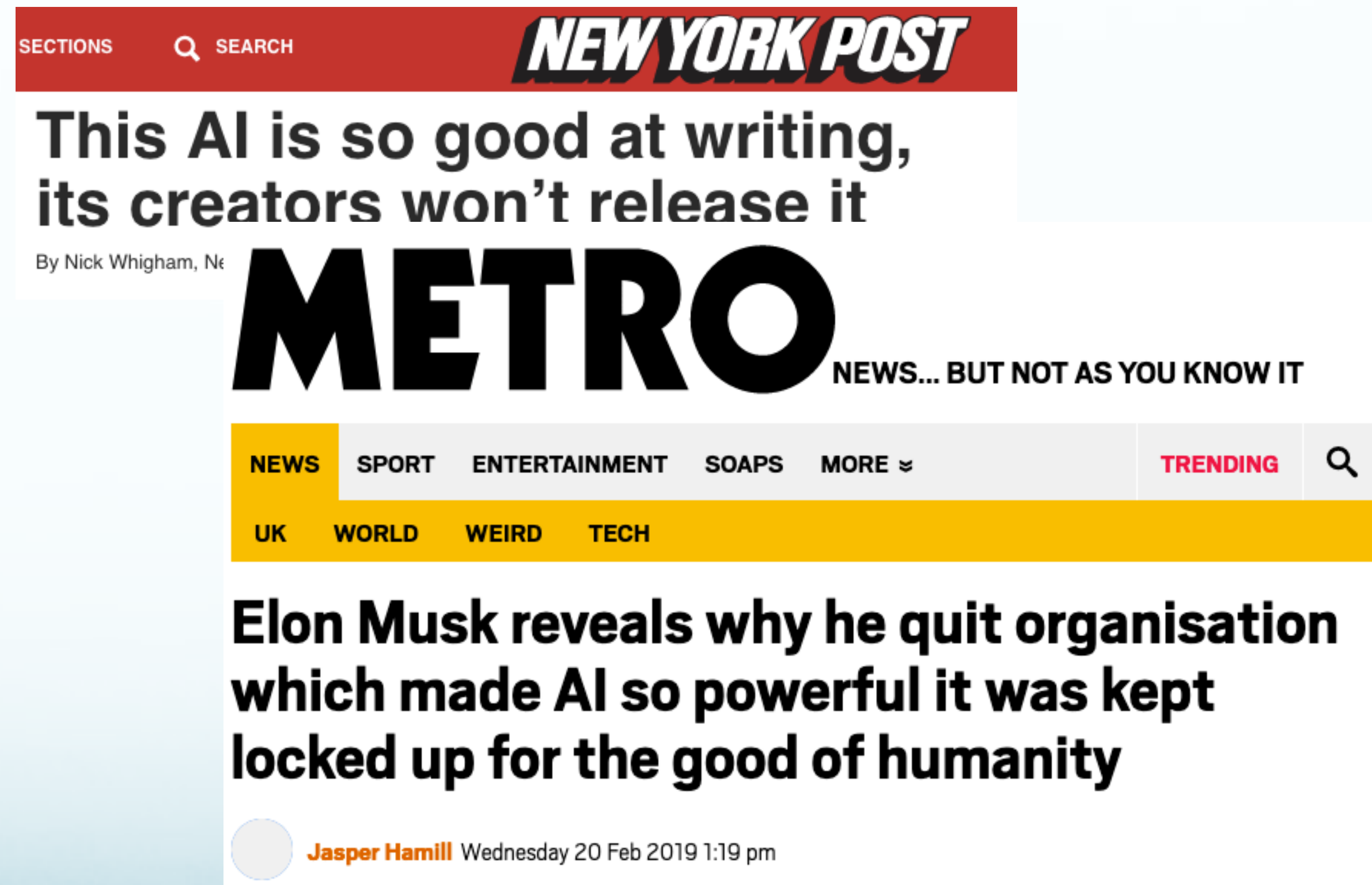
- Headlines:



OpenAI: Language Models, SciComm, and Ethics

- [Original Post](#) - [[code](#)] [[tech report](#)]

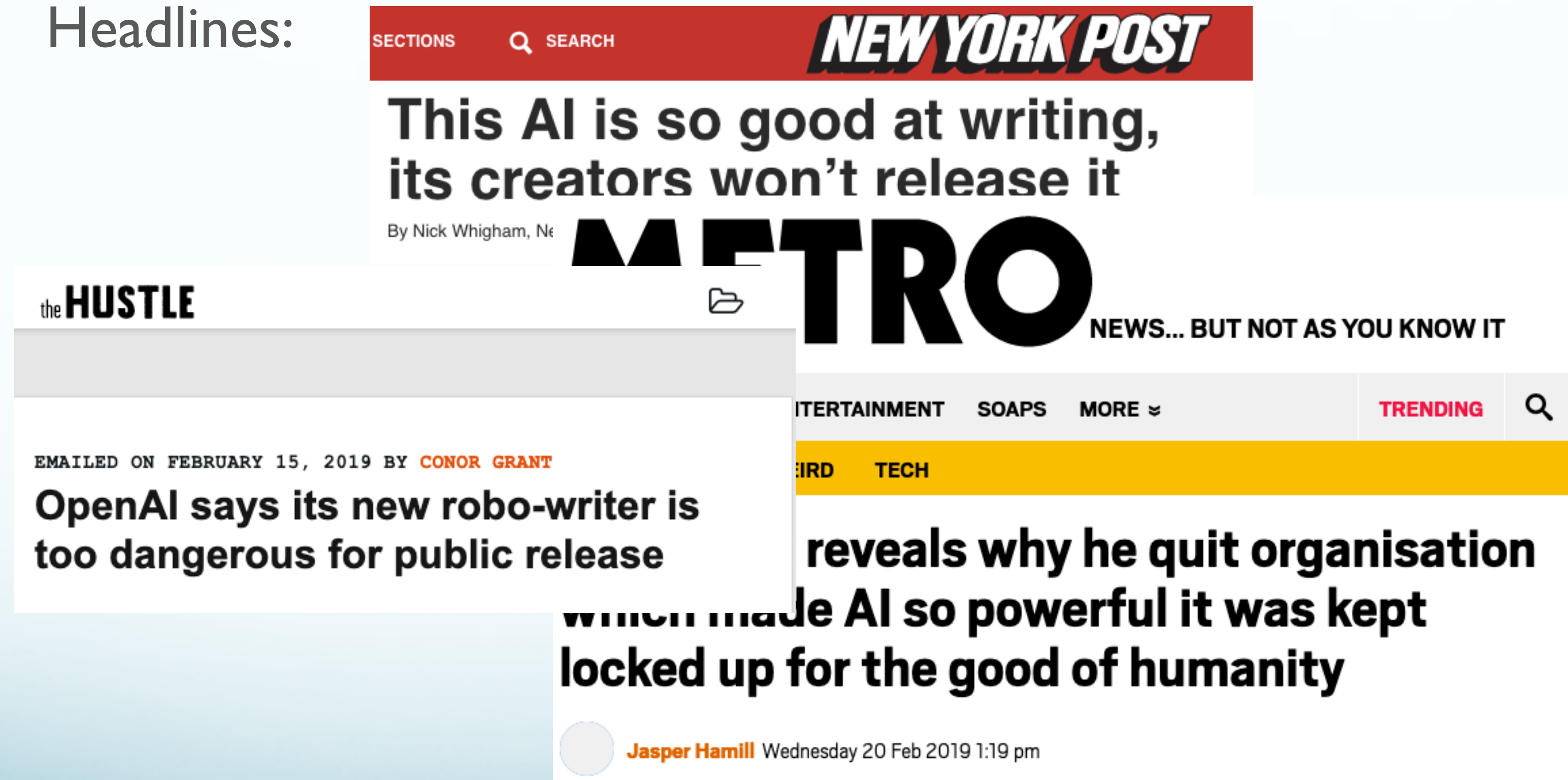
- Headlines:



OpenAI: Language Models, SciComm, and Ethics

- [Original Post](#) - [[code](#)] [[tech report](#)]

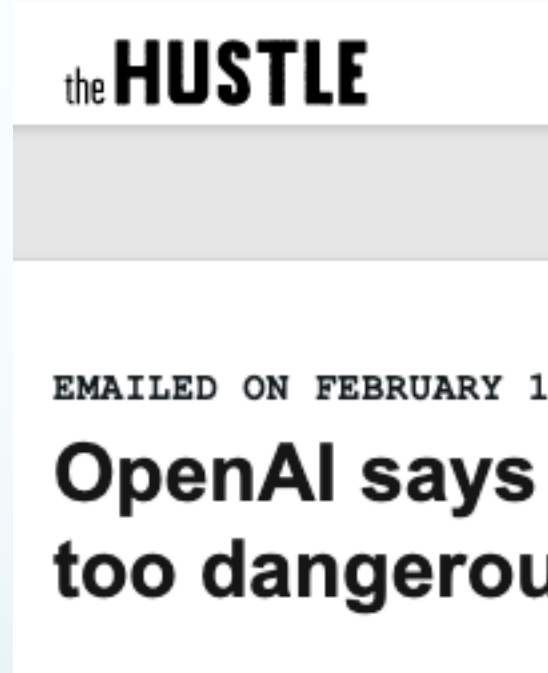
- Headlines:



OpenAI: Language Models, SciComm, and Ethics

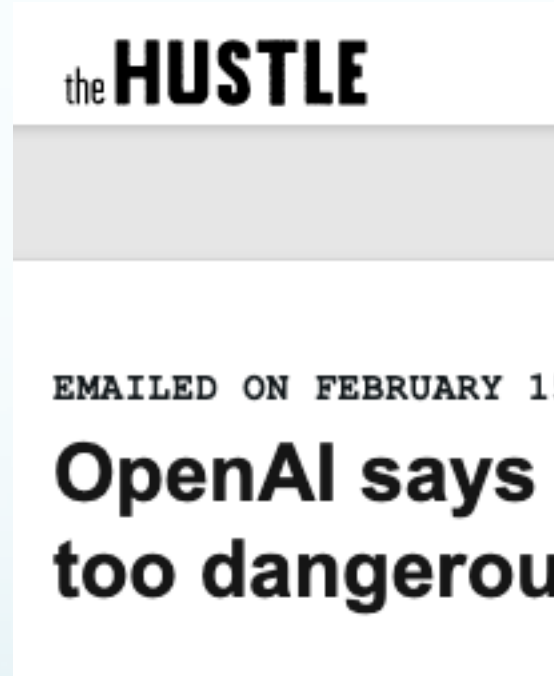
- [Original Post](#) - [[code](#)] [[tech report](#)]

- Headlines:



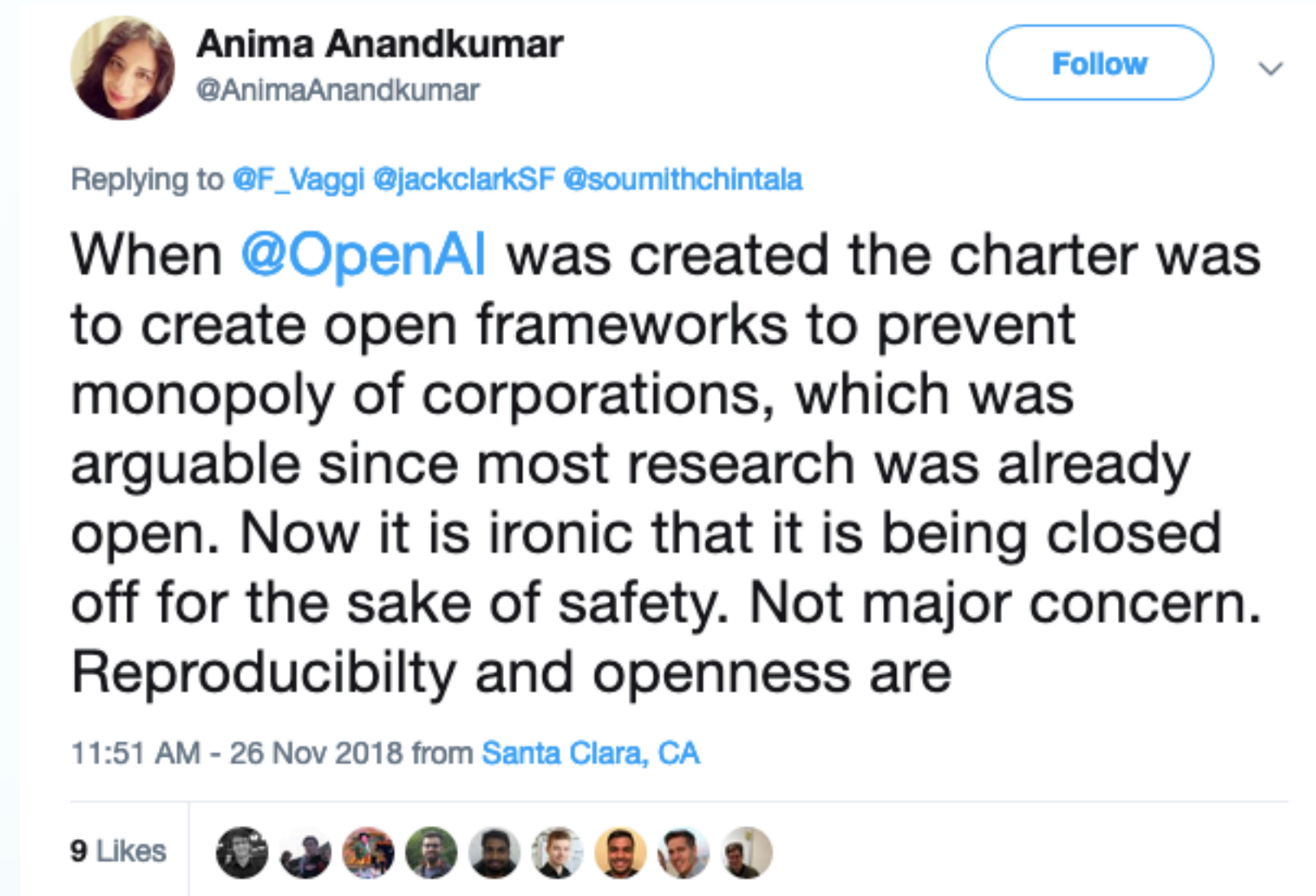
OpenAI: Language Models, SciCo

- [Original Post](#) - [\[code\]](#) [\[tech report\]](#)
- Headlines:



OpenAI: Language Models, SciComm, and Ethics

- Academic Responses:



OpenAI: Language Models, SciComm, and Ethics

- Academic accounts:
 - Anima Anandkumar — [An open and shut case on OpenAI](#)
 - Zachary Lipton — [OpenAI Trains Language Model, Mass Hysteria Ensues](#)
 - Robert Munro — [Should I Open-Source My Model?](#)
 - Ryan Lowe — [OpenAI's GPT-2: the model, the hype, and the controversy](#)

OpenAI:

Language Models, SciComm, and Ethics

- Sample story, from their blog post:

System Prompt (human-written)

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

Model Completion
(machine-written, 10 tries)

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved. Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns.

While examining these bizarre creatures the scientists discovered that the creatures also spoke some fairly regular English. Pérez stated, "We can see, for example, that they have a common 'language,' something like a dialect or dialectic."

OpenAI:

Language Models, SciComm, and Ethics

- Sample, from [Cyril Diagne](#):

System Prompt (human-written)

"You see a runaway trolley moving toward five tied-up (or otherwise incapacitated) people lying on the tracks. You are standing next to a lever that controls a switch. If you pull the lever, the trolley will be redirected onto a side track, and the five people on the main track will be saved. However, there is a single person lying on the side track. You have two options:

1. Do nothing and allow the trolley to kill the five people on the main track.
2. Pull the lever, diverting the trolley onto the side track where it will kill one person. Which is the more ethical option?"

OpenAI:

Language Models, SciComm, and Ethics

- Sample, from [Cyril Diagne](#):

Model
(Temperature: 0.8)

You can always push the lever to kill the five people on the main track and then continue on your own,

The question is how you could prevent this. How do you know when the trolley is on its way to kill someone' Or to save the five people on the main track from drowning? The most obvious way to prevent this is to place a large, open container in the middle of the side track. It will contain a small stack of food and a large food bowl. If you open it and do not find anything, you should continue to open it;

One might suggest, "There are two ways to prevent this:

- 1.) You can keep the food in a small, open container and the container will be moved to a side track.
- 2.) You can open the containers in a small container in the middle of the track. This would prevent the trolley from killing the five people on the main track.

We have seen a number of options. The most obvious means of stopping a trolley is to put a container inside the front of the trolley. This is also one of the more humane ways to prevent drowning. If you open the container and do not find anything, you should not keep it in the front of the trolley.

OpenAI:

Language Models, SciComm, and Ethics

- Sample, from [Cyril Diagne](#):

Model (Temperature: 0.8) You can save the five people on the main track, and the others will be safe, too. Or you can kill the five people on the main track.

Do nothing and allow the trolley to kill the five people on the main track,

Do not pull the lever. The trolley will end up in the trolley room, where the trolley team will be in the trolley room and will be able to get to the place where they are supposed to stop. In this situation, you do not have to wait for the trolley to pass the line where you are supposed to stop; just pull the lever.

The trolley team will be in the trolley room, where the trolley team will be in the trolley room and will be able to get to the place where they are supposed to stop. In this case, you do not have to wait for the trolley to pass the line where you are supposed to stop; just pull the lever. Do not give the trolley the last word, but allow for a delay in the trolley. The team that is supposed to stop the trolley will be in the trolley room, and they will be able to finish the trolley.

The trolley team that is supposed to stop the trolley will be in the trolley room, and they will be able to finish the trolley. Do not pull the trolley. The trolley team will be in the trolley room, and the trolley team will be able to move to the exit. It is best to go to the trolley room immediately, and then go to the exit.

OpenAI:

Language Models, SciComm, and Ethics

- Sample, from [Cyril Diagne](#):

Model (Temperature: 0.2) The most common way to kill a runaway trolley is to pull the lever. However, if you pull the lever, the trolley will be redirected onto a side track, and the five people on the main track will be saved. However, there is a single person lying on the side track, You have two options:

Do nothing and allow the trolley to kill the five people on the main track. Pull the lever, diverting the trolley onto the side track where it will kill one person. Which is the more ethical option? Do you have to stop the trolley? If you do, you will be killed.

If you do, you will be killed. If you do not, you will be killed.

If you do not, you will be killed.

If you do not, you will be killed.

If you do not, you will be killed.

If you do not, you will be killed.

If you do not, you will be killed.

If you do not, you will be killed.

If you do not, you will be killed.

If you do not, you will be killed.

If you do not, you will be killed.

OpenAI: Language Models, SciComm, and Ethics

- Important points:
 - No new algorithmic improvement. This is just using much more data.
 - The paper is not peer-reviewed.

OpenAI: Language Models, SciComm, and Ethics

- Good practices:
 - The authors note that they are “meta-cherry-picking” examples
 - Also note from how many they sampled 👍
 - Authors released the code, and a model trained on a smaller data set
- Bad practices:
 - Journalists invited to interact with the large model system, but not academics
 - Don't see any effort at a lay explanation of what the system is really doing?

OpenAI: Language Models, SciComm, and Ethics

- Questions:
 - Is withholding the larger model the right thing to do?
 - (Keeping in mind that language generation is not the only task the model can perform)
 - What about the interaction with the press?
 - (Keeping in mind that they are a nonprofit that is seeking funding...)

Bias: How it Arises in Data Sets

Two Big Concepts

- Observations vs. Targets:
 - What are you **sampling** vs. what is your **prediction target**?
 - If you knew what the outcome would be, you wouldn't need to build a model.
 - Models sample **observations** that are **proxies** for the **latent variable**.

Two Big Concepts

- Observations vs. Targets:
 - What are you **sampling** vs. what is your **prediction target**?
 - If you knew what the outcome would be, you wouldn't need to build a model.
 - Models sample **observations** that are **proxies** for the **latent variable**.
- What does your sample look like \leftrightarrow Where did your data come from?
 - Knowing your **data provenance** is important for identifying **confounding variables**

Language as a Proxy

Language as a Proxy

- Language ***is a proxy***:
 - For the concepts that humans are trying to communicate
 - ...***which is a proxy*** for the actual state of the world.
 - The “black sheep problem” ([Daumé, 2016](#))
 - ...also remember [Sapir-Whorf](#)

Data Provenance

Google Translate

Text Documents

DETECT LANGUAGE **TURKISH** GERMAN ↕ **ENGLISH** GER

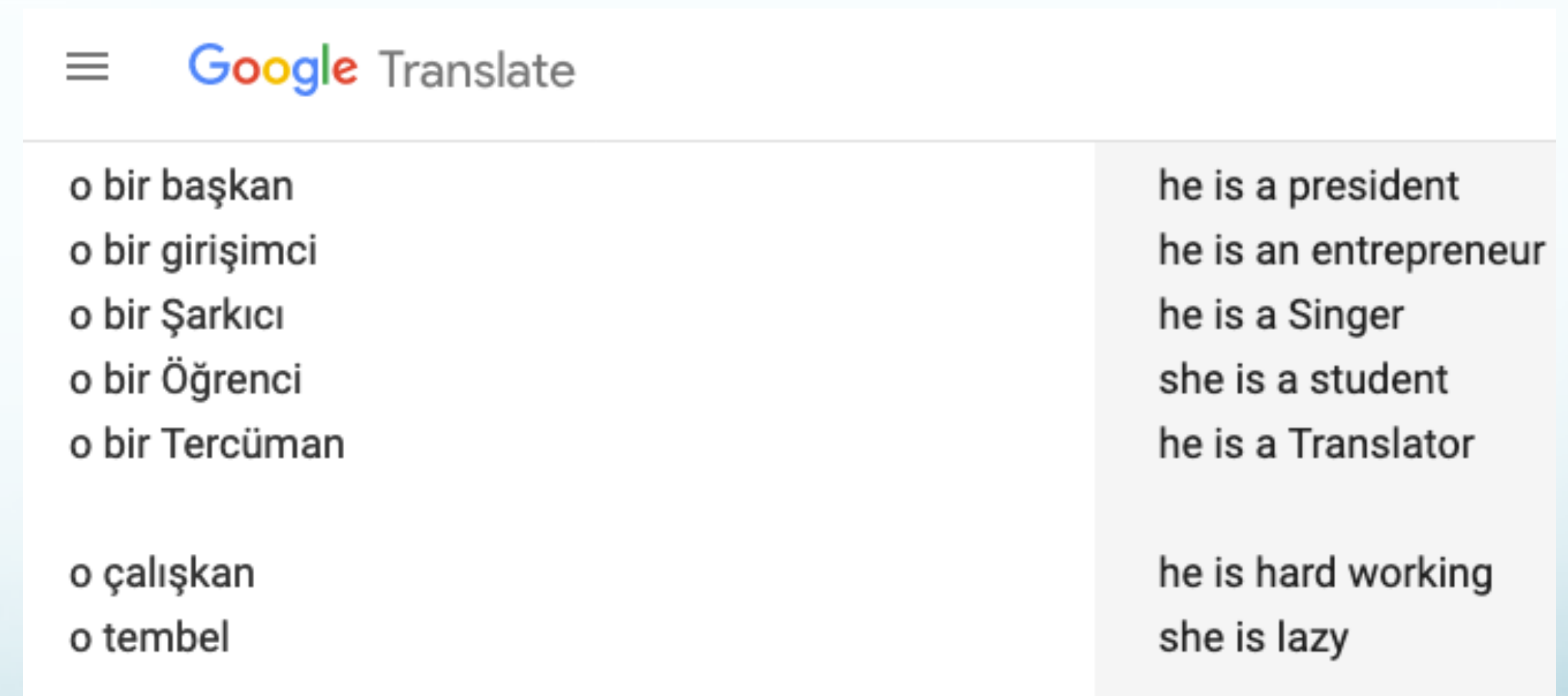
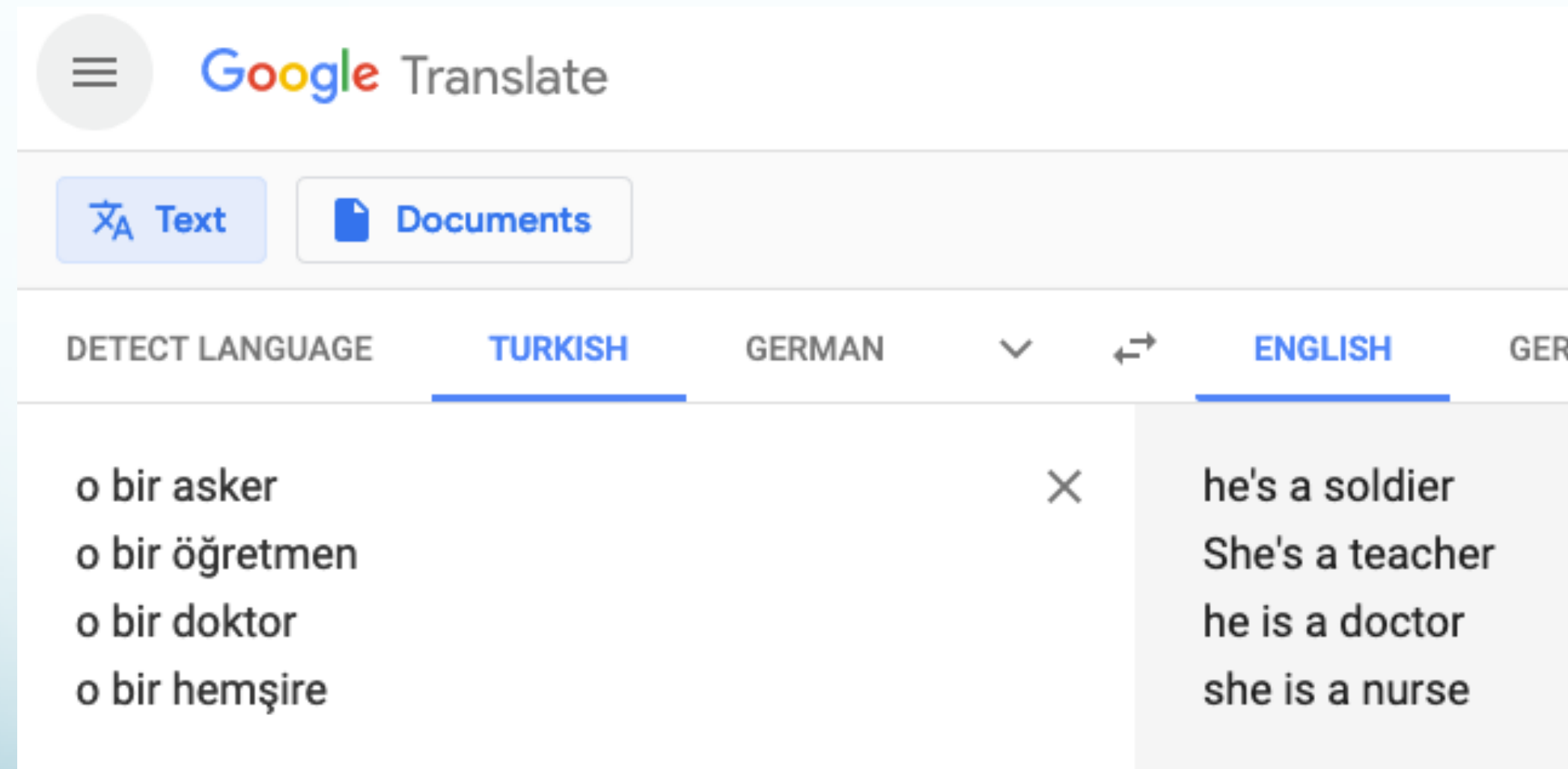
o bir asker	×	he's a soldier
o bir öğretmen		She's a teacher
o bir doktor		he is a doctor
o bir hemşire		she is a nurse

Google Translate

o bir başkan	he is a president
o bir girişimci	he is an entrepreneur
o bir Şarkıcı	he is a Singer
o bir Öğrenci	she is a student
o bir Tercüman	he is a Translator
o çalışan	he is hard working
o tembel	she is lazy

Data Provenance

- Where did your data come from?
 - (Language does not exist in a sociological vacuum)



Data Provenance

The screenshot shows the Google Translate interface with the source language set to Turkish and the target language to English. The input text is "o bir asker". The output shows two gender-specific translations: "she is a soldier (feminine)" and "he is a soldier (masculine)". A notification banner at the bottom right highlights this as a "NEW FEATURE" and explains that Google Translate now displays gender-specific translations for short phrases in this language pair.

DETECT LANGUAGE **TURKISH** GE **ENGLISH** GERMAN SPANISH

o bir asker

Translations are gender-specific. [LEARN MORE](#)

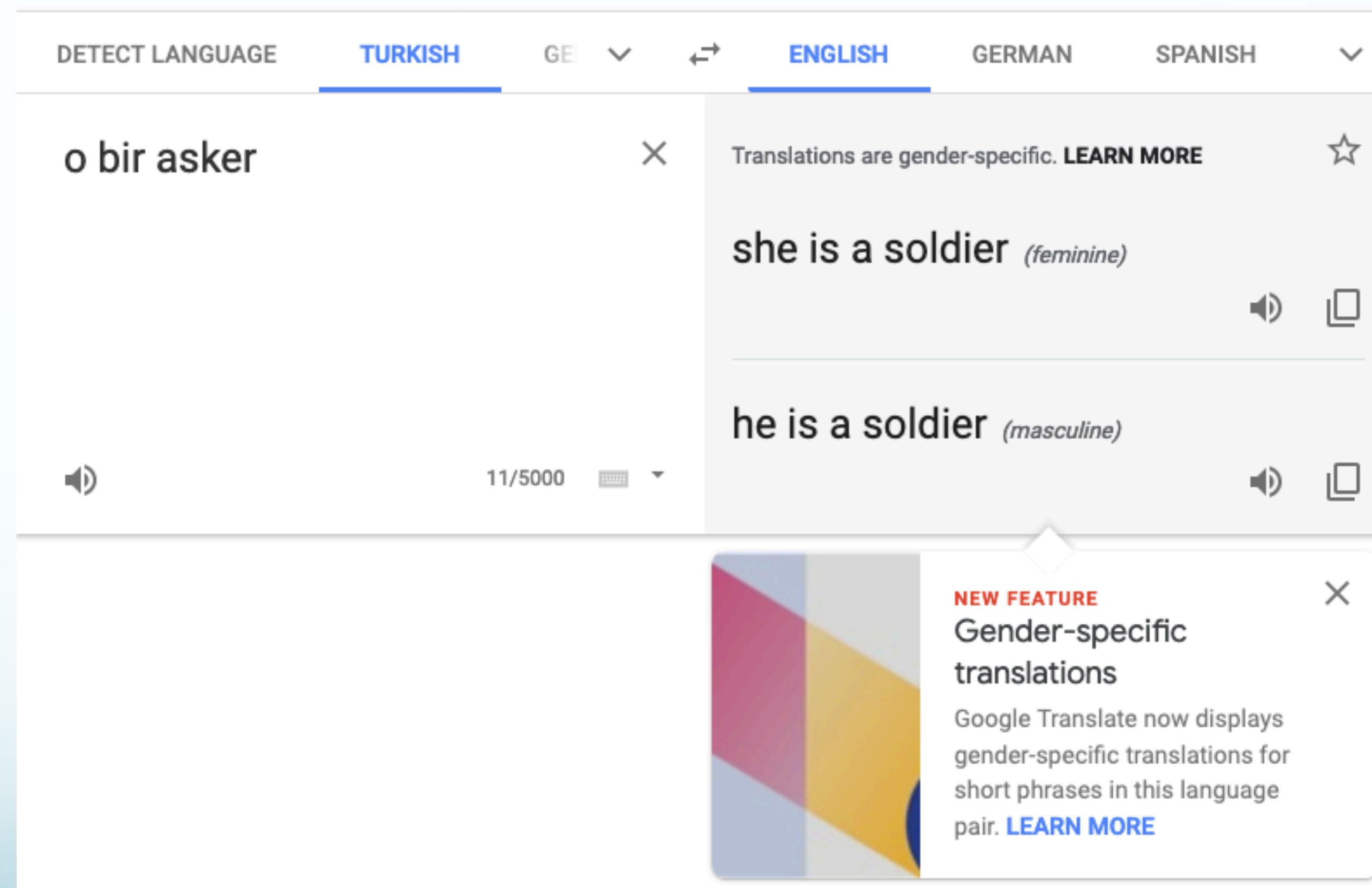
she is a soldier *(feminine)*

he is a soldier *(masculine)*

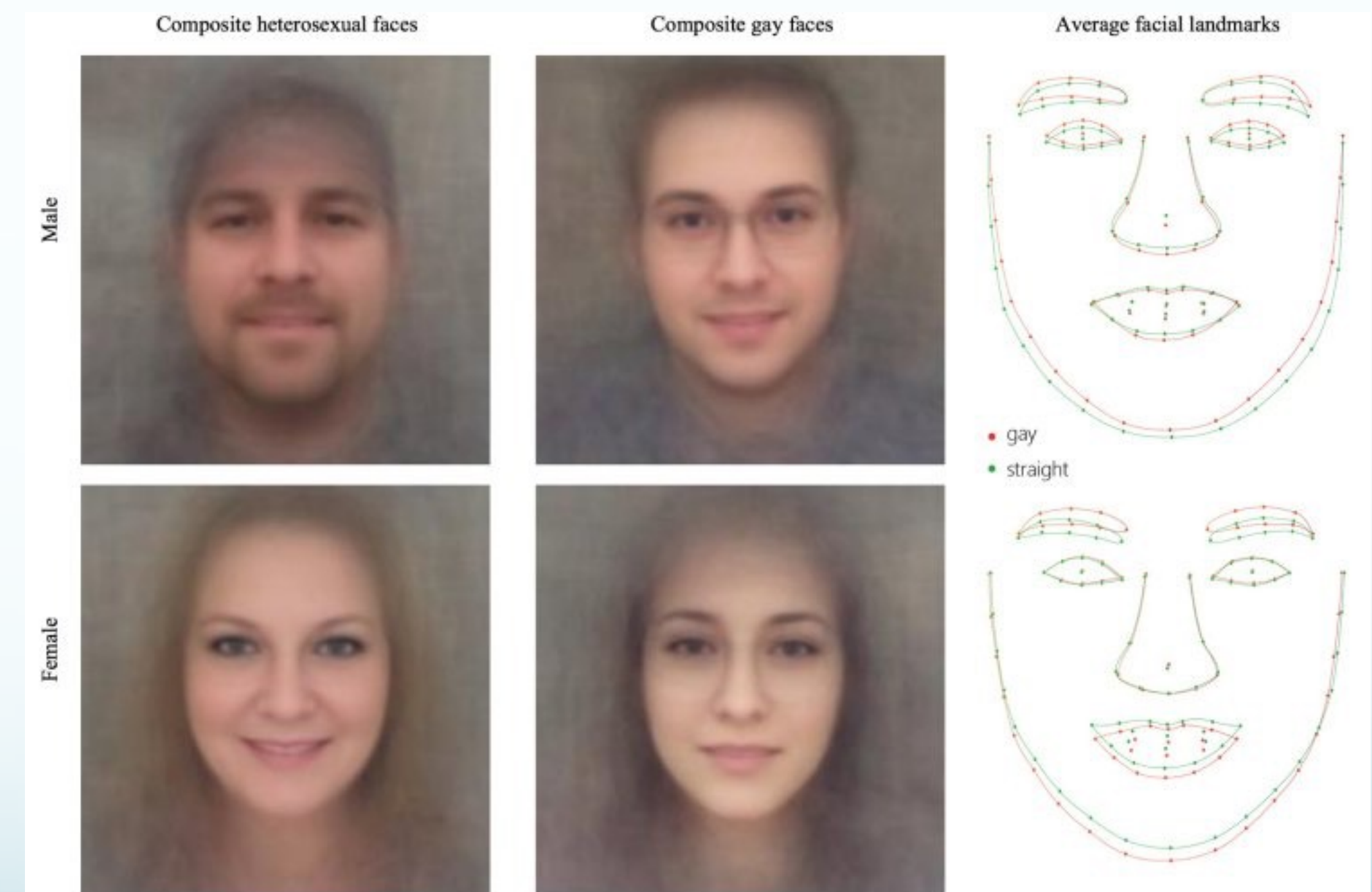
NEW FEATURE
Gender-specific translations
Google Translate now displays gender-specific translations for short phrases in this language pair. [LEARN MORE](#)

Data Provenance

- 2019 Update: when using shorter phrases:

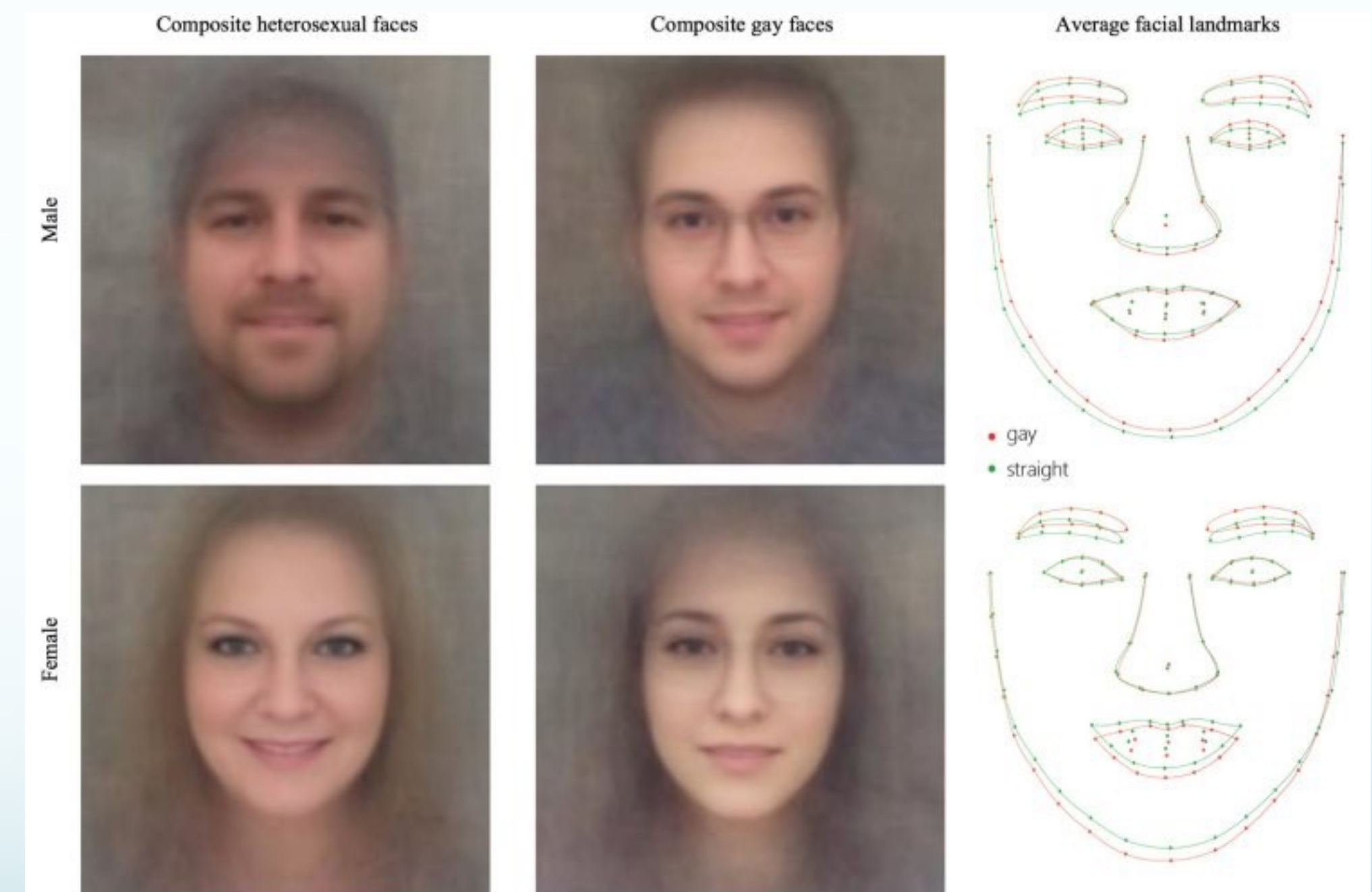


Data Provenance



Data Provenance

- Not NLP, but again, consider social cues:
 - [Wang & Kosinski \(2018\)](#)
 - Data came from a dating website
 - Only white faces
 - Gender, Sexual Orientation:
 - Both are socially performed



Gendering Chatbots

- Also, Sapir-Whorf!