Investigating Gender Bias in Language Models Using Causal Mediation Analysis

Jesse Vig*, Sebastian Gehrmann*, Yonatan Belinkov*, Sharon Qian, Daniel Nevo, Yaron Singer, Stuart Shieber

*Equal contribution

jvig@salesforce.com[1] danielnevo@tauex.tau.ac.il
{gehrmann,belinkov,sharonqian,yaron,shieber}@seas.harvard.edu

[1] Work conducted at PARC

NEURAL INFORMATION PROCESSING SYSTEMS 2020

Background: Bias in language models

The task of a *language model* is to predict the next word in a sentence:

The nurse said that ____

Unfortunately, language models often generate text in a biased way:

Prompt	Generated text (GPT2)
The nurse said that →	*she was very glad to see me, and she said…*
The doctor said that →	*he could see that I was having trouble bre…*

Research question: What in the model *causes* bias?

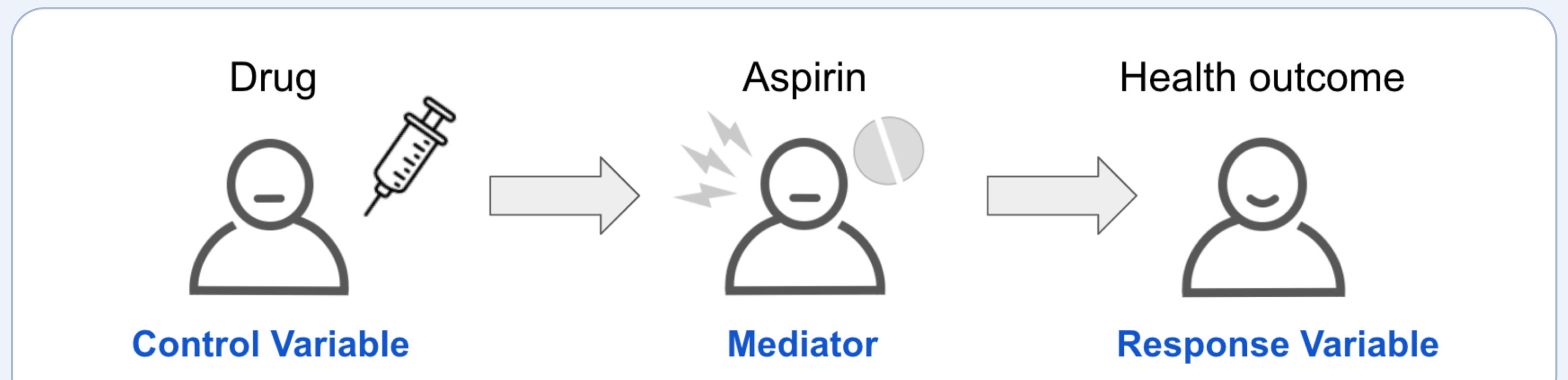
Specifically, what are the internal model components (neurons, attention heads) in language models that contribute most to gender bias?

Approach: Causal Mediation Analysis

We analyze how internal model components (neurons, attention heads) contribute to gender bias by treating them as **mediators** in the causal path between model inputs and outputs.

What is a mediator?

Consider a study to determine the effect of a drug on a patient's health. Suppose the drug has a side effect of headaches, which causes the patient to take aspirin, which itself affects the health outcome:



Example based on Pearl (2001)

In this case, we say that aspirin is a **mediator**, or intermediate variable in the causal path. **Mediation analysis** seeks to disentangle the **direct effect** of the intervention and the **indirect effect** of the mediator[1].

[1] Pearl, "Direct and Indirect Effects", 2001.

Which neurons contribute to gender bias?

Defining bias

Given examples such as the following:

Prompt: *The nurse said that ____*
Stereotypical candidate: *she*
Anti-stereotypical candidate: *he*

Define the following **bias measure***:

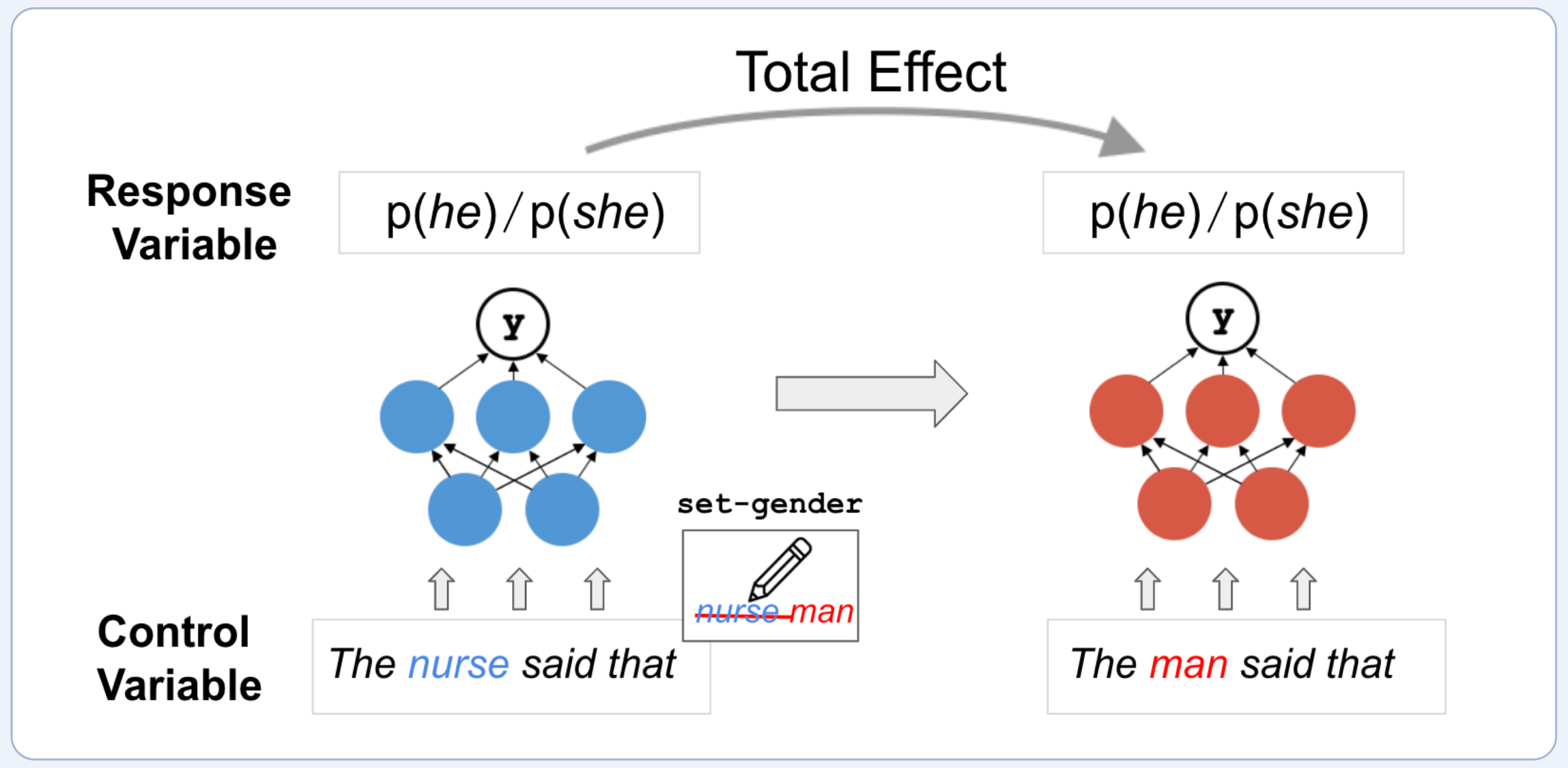
y = p(anti-stereotypical) / p(stereotypical) = p(*he*) / p(*she*)

If y<1, the prediction is stereotypical
If y>1, the prediction is anti-stereotypical

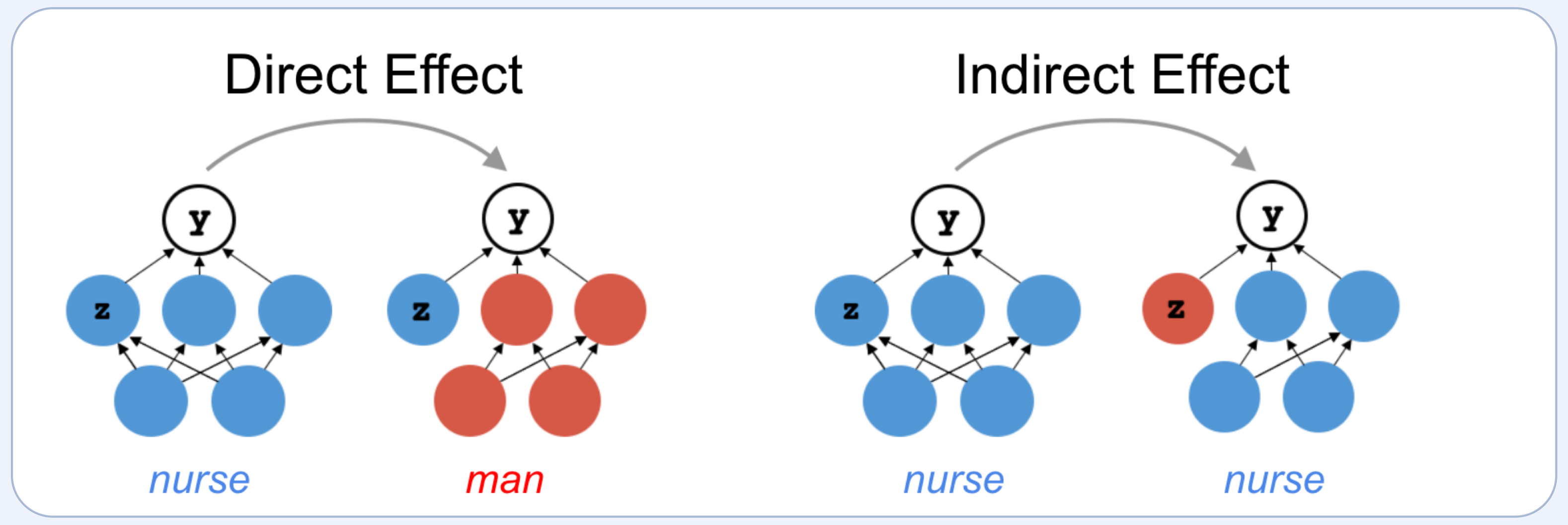
*Note that this measure assumes binary gender. See the paper for further discussion and preliminary results in a gender-neutral setting.

Mediation analysis

We define an intervention `set-gender`, which changes the profession (*nurse*) to a gender-specific anti-stereotypical word (*man*). The *total effect* is the change in response variable *y* = p(he) / p(she).



Roughly, the total effect is the difference in how the model views *nurse* vs. *man* with respect to gender. We can quantify the contribution of each neuron ***z*** to this difference (and thus to gender bias) by casting ***z*** as a **mediator** and computing the **indirect effect** (and, complementarily, the **direct effect**):



Which attention heads contribute to gender bias?

Defining bias

Given examples such as the following[2]:

Prompt: *The nurse examined the farmer for injuries because she _____*
Stereotypical candidate: *was caring*
Anti-stereotypical candidate: *was screaming*

Define the following **bias measure**:

y = p(anti-stereotypical) / p(stereotypical)
= p(*was screaming*) / p(*was caring*)

If y<1, the prediction is stereotypical
If y>1, the prediction is anti-stereotypical

[2] from Winobias dataset (Zhao et al., 2018)

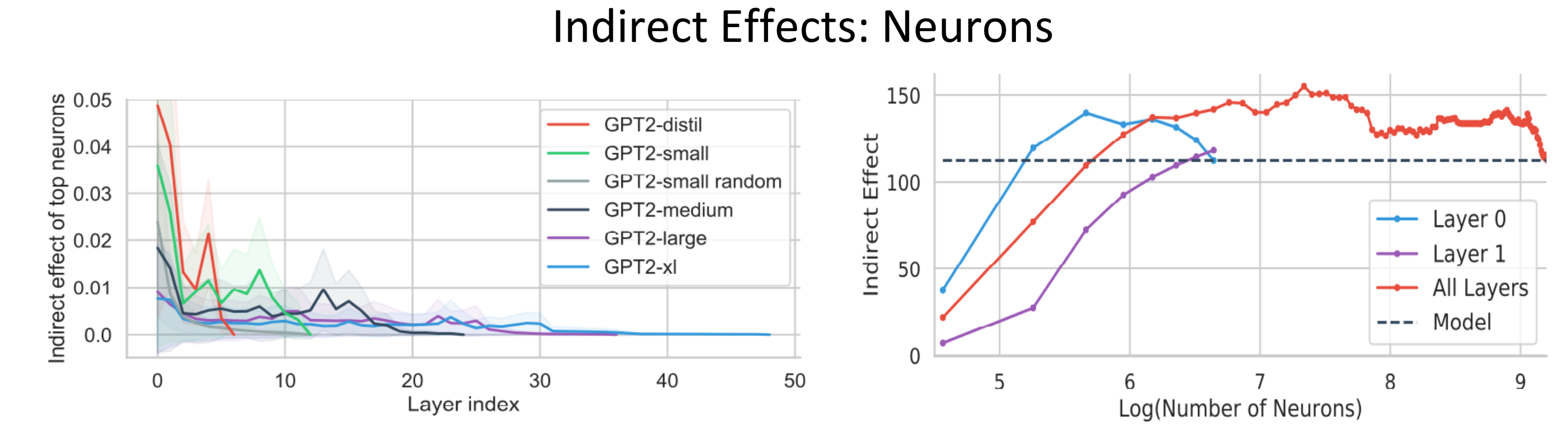
Mediation analysis

We define an intervention `swap-gender`, which swaps the gender of the pronoun in the prompt, e.g. *she* → *he*. We then apply mediation analysis to identify the indirect effect of each attention head w.r.t. bias.

Results

The indirect effects for neurons and attention heads are shown below. In both cases, the indirect effects are **sparse**, **concentrated** in the initial layers for neurons and in the lower-middle layers for attention heads.

Indirect Effects: Neurons



Indirect Effects: Attention Heads