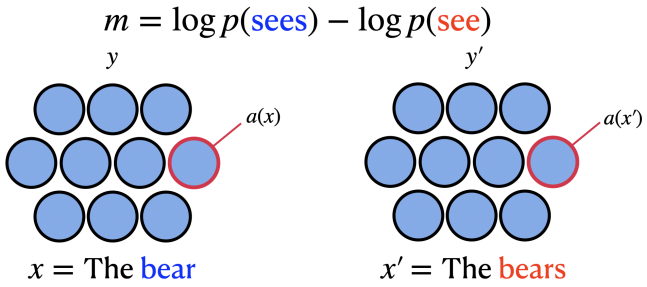
Indirect Effects

You are given the following neural network:



You are also given two prompts, x and x', which differ in grammatical number. Let $a(x)$ and $a(x')$ denote activations of a given neuron in the network given x or x', respectively. We measure the model's output by measuring the log-probability difference between two possible next tokens:

$$y = \log p(\text{sees}) - \log p(\text{see}).$$

To measure the influence of a, we can perform **activation patching**, replacing its activation with some counterfactual activation while holding all else in the model fixed.

Run	**Input**	**Activation**	y
1	x	$a(x)$	2
2	x'	$a(x')$	6
3	x	$a(x')$ (patched)	5
4	x'	$a(x)$ (patched)	4

Problem (1A): Trace and calculate

(a) Identify the run index where activation $a(x')$ is patched into the x run. Also identify the original and counterfactual runs that we would need to perform *before* we can perform this patching run.

(b) Calculate the total effect and indirect effect:

$$\text{TE} = y(x') - y(x),$$
$$\text{IE} = y(x, \text{do}(a := a(x'))) - y(x).$$

In 1–3 sentences, what, intuitively, does each metric tell us?

(c) Perform the reverse experiment: insert $a(x)$ into the x' run, and compute the indirect effect relative to the unpatched x' run. Is it the negative of the first indirect effect? Either way, in 1–3 sentences, what does this outcome tell us about the indirect effect?

Now, we will practice interpreting indirect effects in scientific contexts.

Problem (1B): Interpreting the indirect effect

A researcher writes:

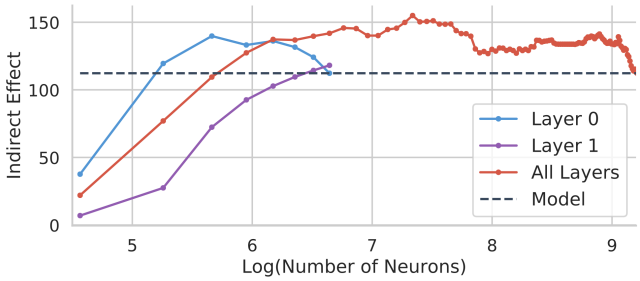
> "This activation contains 75% of the model's gender representation, because IE/TE = 0.75."

(a) What does this ratio really measure? Rewrite the claim so that it is supported by the experiment.

(b) What additional justification would you need to interpret the patch as changing *only gender*? Name something else the patch could change.

(c) Suppose the patched score were 8, while the original and source scores remained 2 and 6. Recalculate IE and IE/TE. Is an indirect effect larger than the total effect impossible? Suggest an explanation.

This figure from Vig et al. [2020] shows the effect of adding progressively more neurons to the set of components we intervene on, and then remeasuring the indirect effect of that set of neurons.



Problem (1C): Interpreting the indirect effect, cont.

Notice how in the above figure, the indirect effect generally increases as we add more neurons to the set up to point. Then, the indirect effect goes down. In 1–2 sentences, what does this decline in the indirect effect tell us?

The Causal Ladder

Here, we will assess whether a given type of evidence justifies the demonstrated claim and the interpretation of the manipulated variable.

Rung	Question
L1: Association	What varies with what in observed runs?
L2: Intervention	What happens when we set a variable?
L3: Counterfactual	Given a factual situation, what would happen under an alternative intervention in that same situation?

Problem (2): Classify the query; audit the evidence

A. For each claim, give its rung and one requirement for supporting it (data, intervention, model, or alignment).

1. A linear probe predicts the carry with 98% accuracy.
2. Zeroing this head reduces addition accuracy by 20 percentage points.
3. For the observed deterministic run on $17 + 25$, replacing a with a' would have changed the answer from 42 to 32.
4. For that same run, removing the model's represented carry would have changed the answer from 42 to 32.
5. Among examples with high activation in this feature, wrong answers are more common.

What extra justification does claim 4 need compared with claim 3?

B. A model originally sampled "42." After patching an activation, a rerun samples "32" using fresh randomness. A researcher concludes: "The original run would have sampled 32 if we had patched it."
What changed besides the activation? What is needed for a counterfactual under the same realized randomness? What could repeated reruns with fresh randomness estimate instead?

Deliverable: Five rung classifications with supporting requirements, followed by short answers to the comparison and randomness questions.

Causal Abstraction

Write two addends as $A = 10A_{10} + A_1$ and $B = 10B_{10} + B_1$. Consider the following high-level addition program:

$$C = \mathbf{1}[A_1 + B_1 \geq 10], \qquad O_1 = (A_1 + B_1) \bmod 10,$$
$$O_{10} = A_{10} + B_{10} + C, \qquad \text{Output} = 10O_{10} + O_1.$$

We hypothesize that a chosen network component implements C. Copy the source's carry into the base program, keeping all other base inputs fixed. Separately, copy the aligned network component from the source run into the base run.

Problem (3A): Does the patch behave like a carry?

Cover the final column while filling in the program's predicted outputs.

Base	Source	Base C	Source C	Predicted output	Network output
$17 + 25$	$30 + 47$	1	0	_____	32
$30 + 47$	$17 + 25$	0	1	_____	87
$17 + 25$	$18 + 24$	1	1	_____	42
$17 + 25$	$11 + 21$	1	0	_____	22

(a) Compute interchange intervention accuracy:

$$\text{IIA} = \frac{\text{number of matching program and network outputs}}{\text{number of tested base–source pairs}}.$$

(b) Why should the first predicted output differ from the source's answer, 77?

(c) What does the third row test, even though the carry does not change?

Deliverable: Four predictions, an IIA value, and two short explanations.

References

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc., 2020.