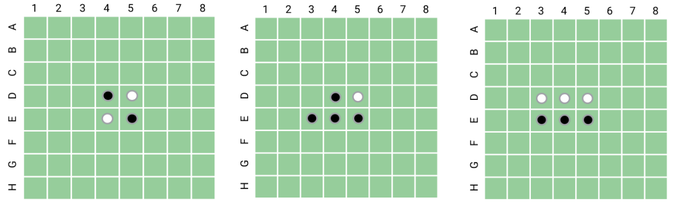
Emergent world representations:
Exploring a sequence model trained on a synthetic task.

Kenneth Li, Aspen Hopkins, David Bau, Fernanda Viegas, Hanspeter Pfister, & Martin Wattenberg

Game of Othello As Toy Model



Legality is nonlinear function of board state and the board state is a nonlinear function of the moves.

State-space complexity (number of unique states): <10^28
Game tree complexity (number of leaves): 10^58

→ E3, D3, C3, … → 32, 26, 18, … → 60 x 512 **Word Embeddings**

Othello-GPT

Trained separate 8-layer, 8 attention-headed GPTs on:

1) World championship dataset (132k games), &
2) Synthetic dataset (20m games)

Evaluation on games in held-out validation set—whether the top prediction is illegal given game rules, e.g. our error:

- *Trained on synthetic dataset:* 0.01%
- *Trained on championship dataset:* 5.17%
- *Untrained:* 93.29%

Probing Othello-GPTs

For each square **s** on the board, can we train a simple classifier f_s, such that $f_s(x_i) = \{\textbf{white}, \textbf{black}, \textbf{empty}\}$ (where x_i represents the value of concept C in the input) reflecting whether **s** is white, black, or empty?

Error rates of linear probe across different layers

	x^1	x^2	x^3	x^4	x^5	x^6	x^7	x^8
Randomized	26.7	27.1	27.6	28.0	28.3	28.5	28.7	28.9
Championship	24.2	23.8	23.7	23.6	23.6	23.7	23.8	24.3
Synthetic	21.9	20.5	20.4	20.6	21.1	21.6	22.2	23.1

Error rates of nonlinear probe across different layers

	x^1	x^2	x^3	x^4	x^5	x^6	x^7	x^8
Randomized	25.5	25.4	25.5	25.8	26.0	26.2	26.2	26.4
Championship	12.8	10.3	9.5	9.4	9.8	10.5	11.4	12.4
Synthetic	11.3	7.5	4.8	3.4	2.4	1.8	1.7	4.6

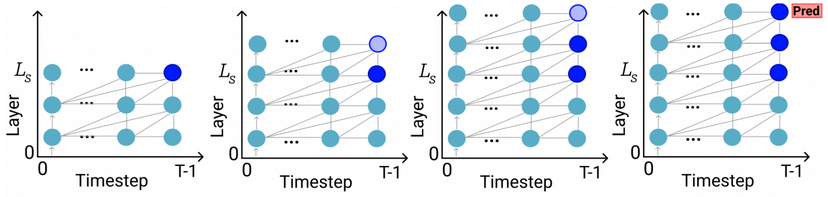
Intervention Technique

An important question: are representations found through probing meaningful & causal? To answer this, the representation must meet the following requirements:

Interpretable: Understandable by a human, and related to human models of the same situation

Controllable: Changing the representation changes the output of the system in a predictable way

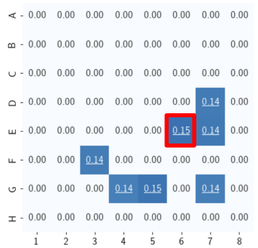
But how can we control the representation?



We can replace original internal representations with a post-intervention representation!

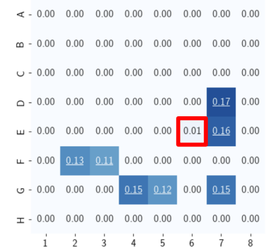
Pause, hack, put-back (& continue).

Logit map ***before*** flipping E3 world representation



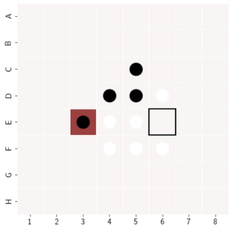
0.15

Logit map ***after*** flipping E3 world representation



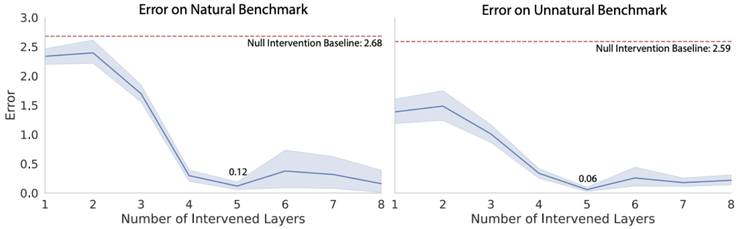
0.01

Updated Board



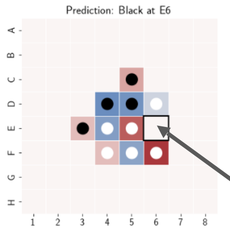
0.14

: E6, the square we attribute



Latent Saliency Maps: Attribution Via Intervention

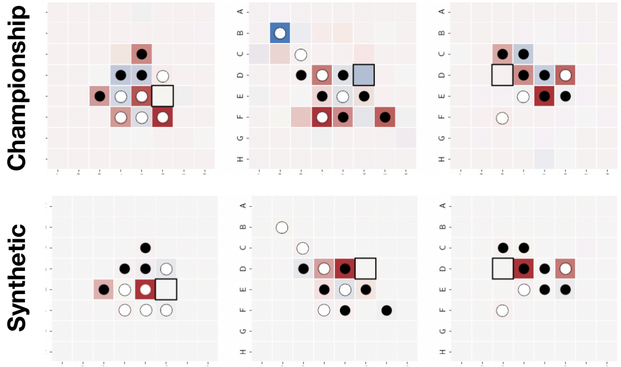
By holding the world model of the colored tile as it is, Othello-GPT thinks it is…



Negatively **Positively**

…contributing to the prediction of the next step.

Enclosed E6: a legal next step we attribute



Harvard John A. Paulson School of Engineering and Applied Sciences
MIT CSAIL
MIT EECS
Northeastern University Khoury College of Computer Sciences