Exploring the Limits of OthelloGPT's Emergent Representations

Julian Baldwin

Overview

- Motivations
- Background
    - Othello-GPT
- Results
- High-Level Takeaways

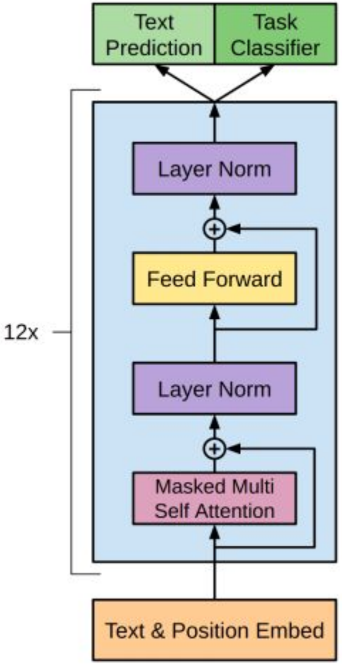
Broad Goals / Motivations

Mechanistic Interpretability: the study of attempting to reverse engineer neural networks to human interpretable algorithms

Current techniques for steering LLM outputs are limited, difficult to make any rigorous guarantees about model behavior

Sub-question: how do transformer models represent concepts?

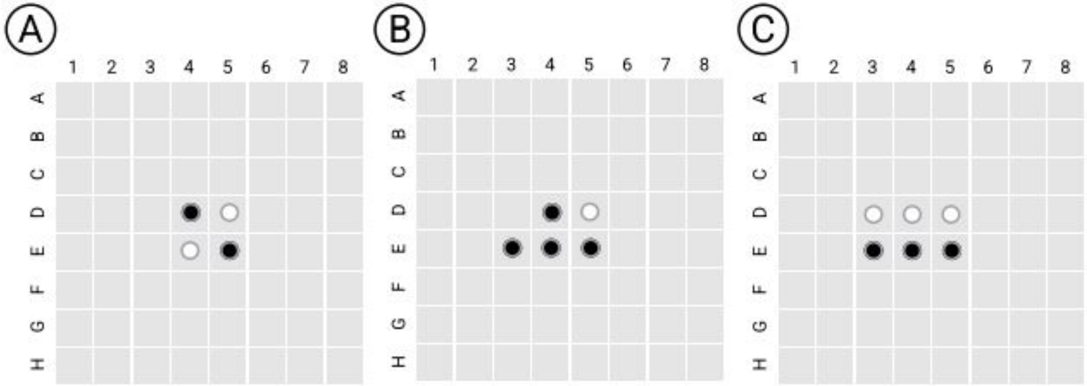
OTHELLO-GPT



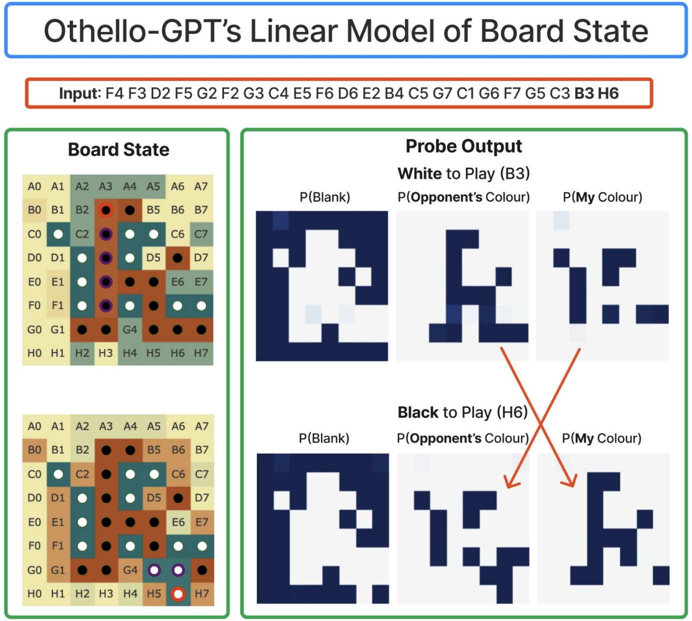
Used the toy task of Othello as a testbed; model is trained on randomly generated games to predict legal moves

Even with no a priori knowledge of the rules or game board, model learns an interpretable representation of the board state

We can extract this representation from the transformer's residual stream using a probe—a classifier trained on model activations



LINEAR REPRESENTATION



Othello-GPT's Linear Model of Board State

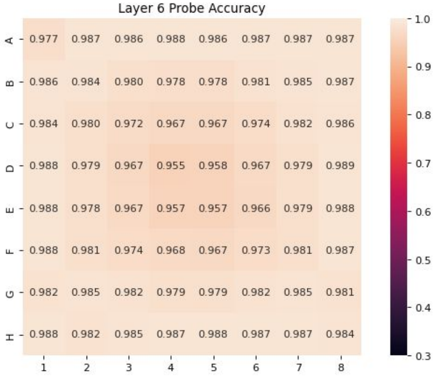
Input: F4 F3 D2 F5 G2 F2 G3 C4 E5 F6 D6 E2 B4 C5 G7 C1 G6 F7 G5 C3 **B3 H6**

Published by Neel Nanda in March 2023

Demonstrates that board state can be recovered with linear probes

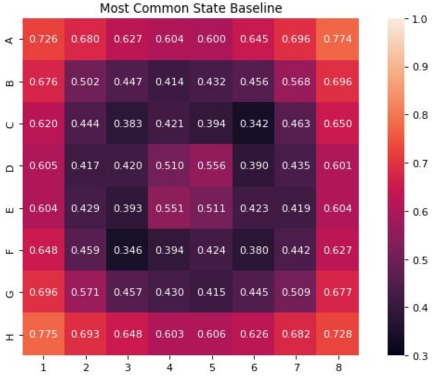
View board as empty/mine/theirs rather than empty/black/white

Recreating Linear Probes



Layer 6 Probe Accuracy

	1	2	3	4	5	6	7	8
A	0.977	0.987	0.986	0.988	0.986	0.987	0.987	0.987
B	0.986	0.984	0.980	0.978	0.978	0.981	0.985	0.987
C	0.984	0.980	0.972	0.967	0.967	0.974	0.982	0.986
D	0.988	0.979	0.967	0.955	0.958	0.967	0.979	0.989
E	0.988	0.978	0.967	0.957	0.957	0.966	0.979	0.988
F	0.988	0.981	0.974	0.968	0.967	0.973	0.981	0.987
G	0.982	0.985	0.982	0.979	0.979	0.982	0.985	0.981
H	0.988	0.982	0.985	0.987	0.988	0.987	0.987	0.984

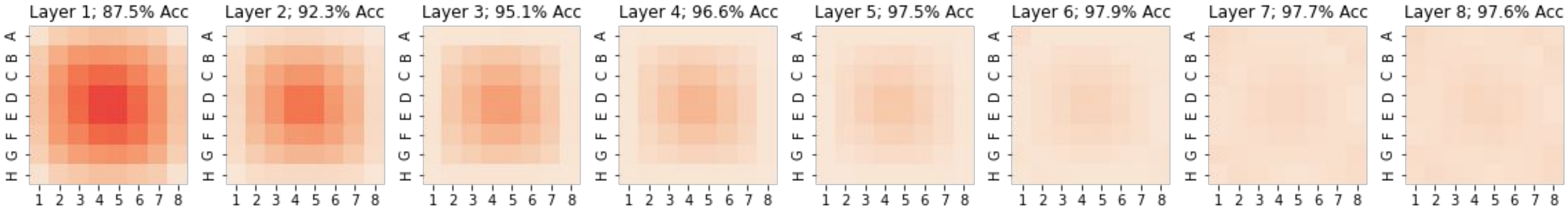


Most Common State Baseline

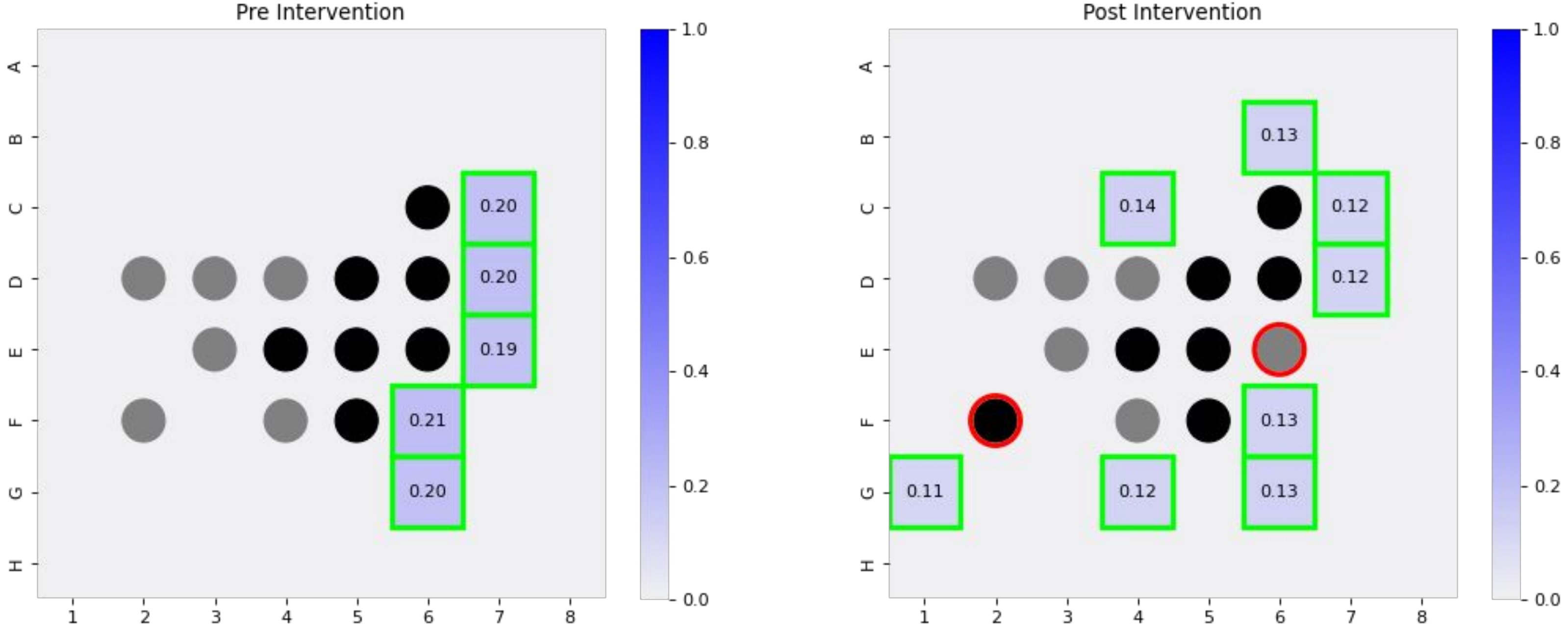
	1	2	3	4	5	6	7	8
A	0.726	0.680	0.627	0.604	0.600	0.645	0.696	0.774
B	0.676	0.502	0.447	0.414	0.432	0.456	0.568	0.696
C	0.620	0.444	0.383	0.421	0.394	0.342	0.463	0.650
D	0.605	0.417	0.420	0.510	0.556	0.390	0.435	0.601
E	0.604	0.429	0.393	0.551	0.511	0.423	0.419	0.604
F	0.648	0.459	0.346	0.394	0.424	0.380	0.442	0.627
G	0.696	0.571	0.457	0.430	0.415	0.445	0.509	0.677
H	0.775	0.693	0.648	0.603	0.606	0.626	0.682	0.728

Starting from the original OthelloGPT codebase, I trained a new model on randomly generated othello sequences

Recreated Neel Nanda's work, achieved high accuracy of board prediction with linear probes

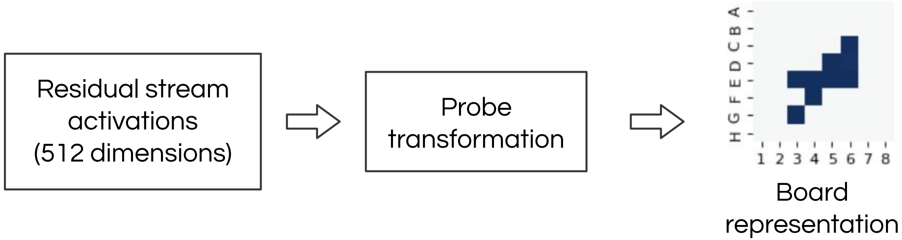


Local Linear Interventions



Flipping the color of two board squares, layer 4 linear intervention

Global Intervention via Probe inverse

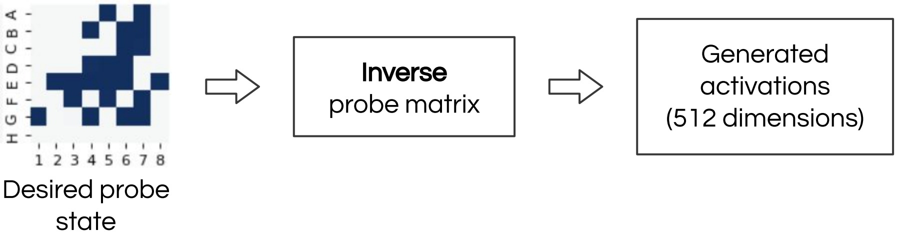


Residual stream activations (512 dimensions)

Probe transformation

Board representation

Usual procedure for extracting board representation from model



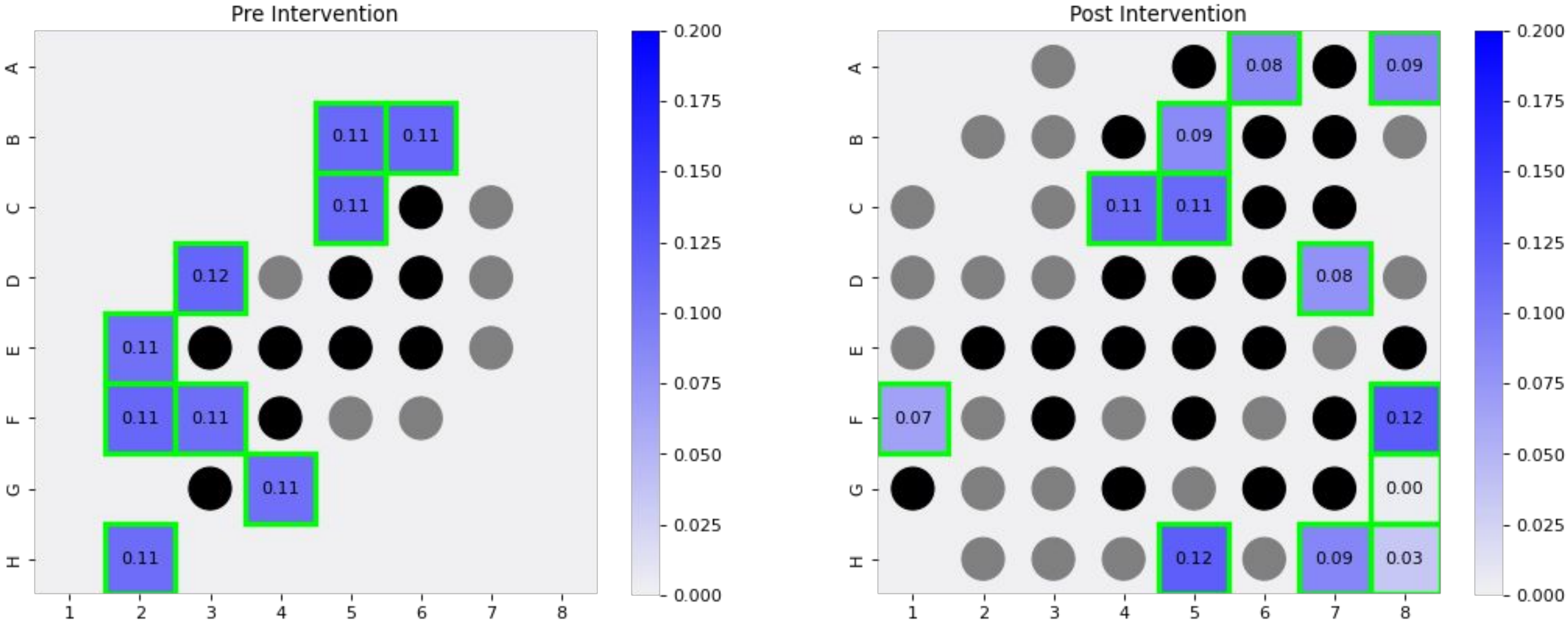
Desired probe state

Inverse probe matrix

Generated activations (512 dimensions)

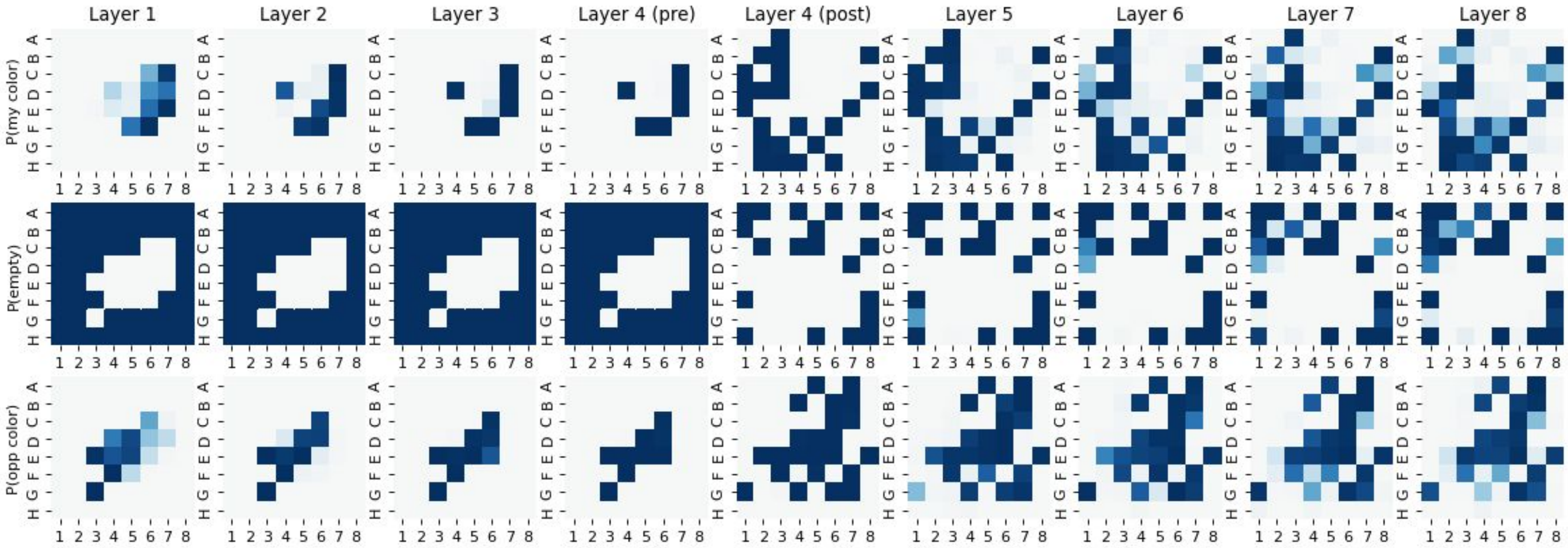
Novel technique for generating activations directly from a desired probe state, which can be inserted into a model during inference

GLOBAL EDITS



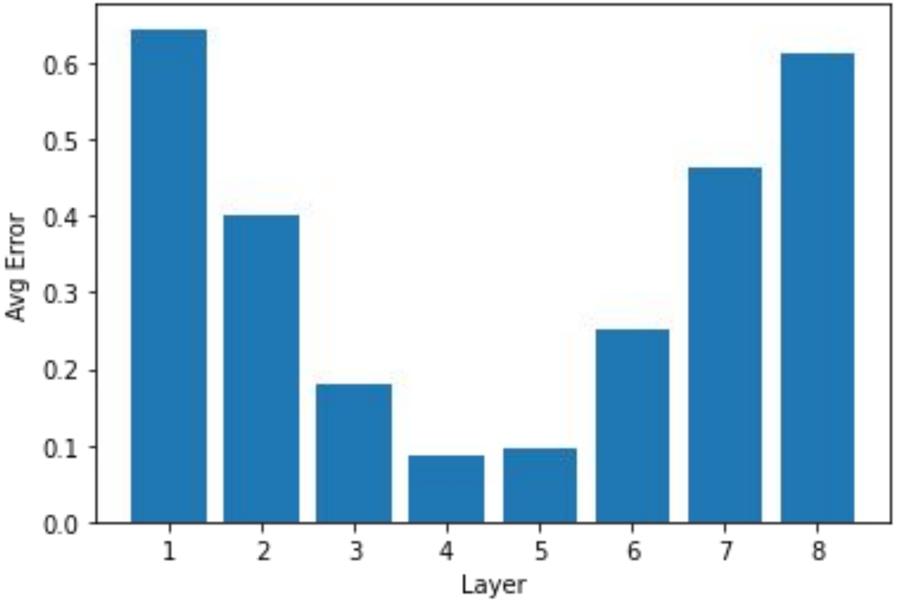
A global intervention: completely overwriting the residual stream after layer 4

Global edits



Board representation visualized across layers during global intervention

Evidence of Board locality

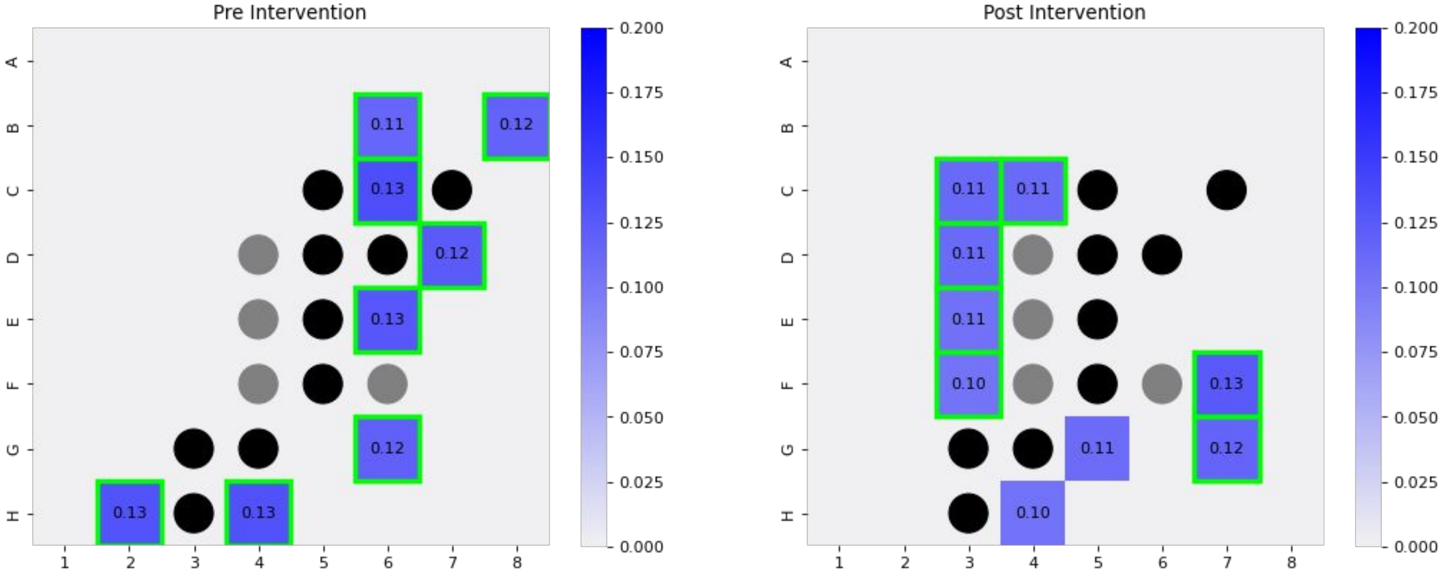


Aggregated results from systemically attempting thousands of global interventions

Error is minimized in layers 4 and 5 where the board has the cleanest representation within the model

Probing Other representations

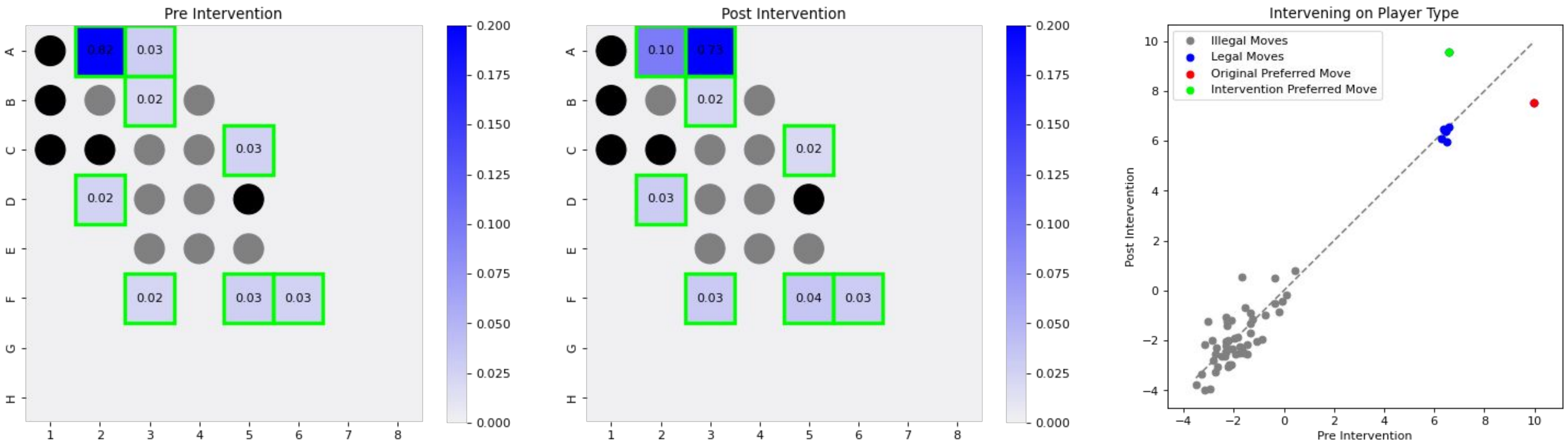
Next color to play: performed successful interventions to shift model between white/black on the same sequence



Intervention on next color from white to black

PROBING OTHER REPRESENTATIONS

Player bias: created new dataset with artificial feature—each game had a bias to move towards one of the corners. Probes were highly accurate but interventions had mixed success



Intervention on player bias from top left to top right, layers 4-7

High-level takeaways

Reinforces linear representation hypothesis

Demonstrates impressive resilience of world models

Probes are an effective tool to extract representations, make causal changes to model without needing more fine-grained analysis, but have fundamental weaknesses in identifying fully disentangled representations

Future work should aim to identify if transformers generally tend towards highly interpretable representations, or just a function of data richness (or some domain specific phenomenon)?

ACKNOWLEDGEMENTS

Huge thanks to **David Reber**, my research mentor for this project, for invaluable advice and guidance throughout the research process

Victor Veitch, for detailed feedback on ideas and access to the DSI cluster so I could train models

Zack, for overall support and running this program

QUESTIONS?