

UNDER WORK IN PROGRESS!! paper is not finished, check back never!!
--

Decodability Is Not Mechanism: Separating What a Network Encodes from What It Uses

Cooper E.

Independent researcher · jcooperkai@gmail.com

ABSTRACT

Interpretability frequently establishes that a network "represents" some quantity by training a probe on its activations and reporting the probe's accuracy. We demonstrate, on a system whose ground truth is known analytically, that decodability and causal mechanism are distinct quantities that can be made to disagree almost completely. A channel decodable from the hidden layer at 100 percent accuracy contributes 0.9998 ± 0.0002 bits to a correlational description of that layer and 0.0006 ± 0.0003 bits to an interventional one -- a separation of roughly 1600 times, stable across five seeds. The interventional description simultaneously recovers the true generating structure at adjusted Rand index 0.9957 ± 0.0054 , where the correlational description reaches 0.1243 ± 0.0305 . Estimating transitions between recovered states reconstructs the generating machine's transition matrix to a maximum absolute error of 0.0154, an object a feature dictionary cannot express. We also report a negative result: selecting the number of states blindly failed under two independent principled criteria, which indicates that a trained network's causal states are approximate rather than exact, and that a state or feature count is undefined without a stated tolerance. This is a controlled demonstration rather than a claim of novelty; the surrounding ideas overlap established work on causal abstraction and on computational mechanics applied to neural networks.

1. The confound

If a variable can be decoded from a network's activations, it is present in them. It does not follow that the network uses it. A probe measures what an external classifier can extract; it does not measure what the downstream computation depends upon. The two coincide often enough that the distinction is easy to lose, and a probe accuracy is easy to report as though it were a mechanistic finding.

The question this note answers is quantitative rather than philosophical: how far apart can the two quantities be driven, on a system where the correct answer is known in advance?

2. A system with known ground truth

The generating process is a unifilar, synchronising source with three causal states known analytically. Unifilarity matters: the emitted symbol determines the successor state, so the process synchronises and a

finite observation window pins the state exactly. Recovery is therefore a well-posed question with a checkable answer.

An earlier version of this experiment used a generic hidden Markov model, which was a mis-specification worth recording. Observing a generic HMM through noisy emissions makes the optimal predictor's causal states *belief* states, and for a generic model those form a continuum. "Recover three clusters" was never the correct target, and a method scored against it would have been penalised for being right.

A network is trained on next-symbol prediction. Its input carries one additional channel of pure noise, amplified so that it dominates the variance of the hidden layer. By construction that channel is fully decodable and causally inert: the target depends only on the state. This is the cleanest possible instance of the confound, built deliberately.

3. Two equivalence relations on the same activations

Correlational. Group activations by proximity in activation space. This is what probing, representational similarity and clustering-based analyses effectively measure.

Interventional. Group activations a and a' together exactly when they induce the same distribution over outputs under intervention:

$$a \sim a' \text{ iff } P(Y | \text{do}(h = a)) = P(Y | \text{do}(h = a'))$$

For a feedforward network the interventional distribution is computed exactly by pushing an activation through the remaining layers, so on this system the relation is evaluated without estimator error. Whatever the two relations disagree about is therefore a property of the definitions, not of an approximation.

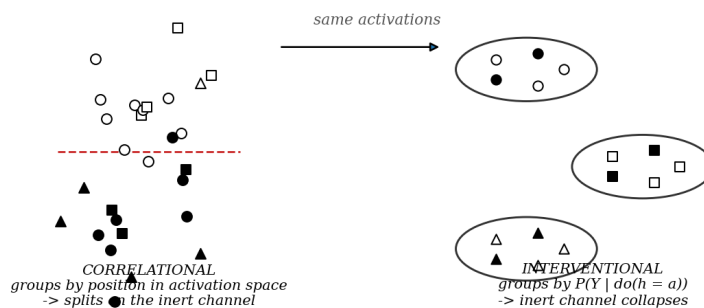


Figure 1. The same activations under two groupings. Marker shape denotes the true causal state; fill denotes the inert channel. Grouping by position in activation space splits along the inert channel, because that channel dominates the variance. Grouping by interventional effect collapses it, because the downstream computation does not depend on it.

4. Result

Measurement	Correlational	Interventional
Mutual information with the inert channel (bits)	0.9998 ± 0.0002	0.0006 ± 0.0003
Adjusted Rand index against true causal states	0.1243 ± 0.0305	0.9957 ± 0.0054

Table 1. Five seeds, mean and standard deviation. The channel is decodable at 100 percent throughout. A fair coin carries exactly one bit, so the correlational grouping captures essentially the entire content of a variable the network never uses, while recovering about a tenth of the structure it does use.

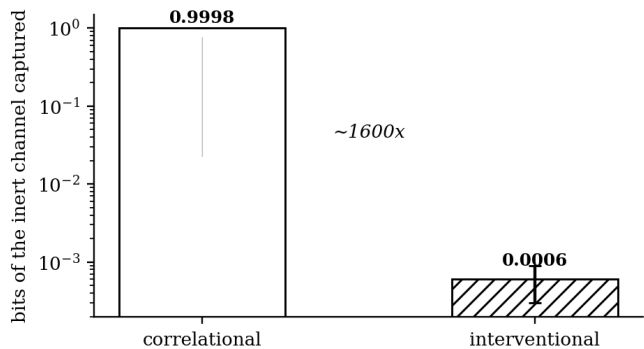


Figure 2. The separation, logarithmic scale. Error bars are standard deviations over five seeds; the effect is far larger than the seed variation.

The mechanism is straightforward once stated. Intervening asks what the network *uses*. Information that is present but unused induces no change in the output distribution, so all activations differing only in that information fall into a single class. The inert channel is therefore quotiented away by construction rather than by tuning.

5. Recovering the algorithm, not only the features

Grouping states is not an account of a computation. A description that stops at features says what the parts are and nothing about how they compose.

Estimating transitions between the recovered states reconstructs the generating machine's transition matrix to a maximum absolute error of 0.0154 and a mean absolute error of 0.0088, read off a trained network with no access to the generator.

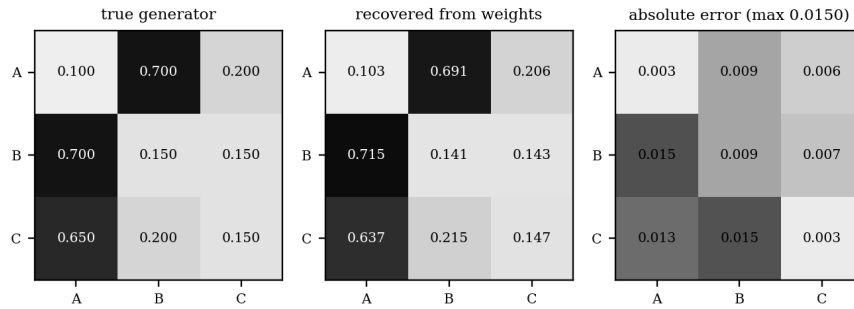


Figure 3. True and recovered transition matrices with absolute error. This object -- states together with the transitions between them -- is what makes the description an algorithm rather than a vocabulary. A feature dictionary has no notion of what follows what and cannot express it.

6. A negative result: the state count is not well defined

Selecting the number of states without being told it failed under two independent principled criteria.

A BIC criterion saturated at the maximum of its search range on every seed, selecting seven when the truth was three. A merge-until-prediction-degrades criterion found no plateau at all: the count slid continuously from about 155 down to 3 as the tolerance widened, with no stable interval.

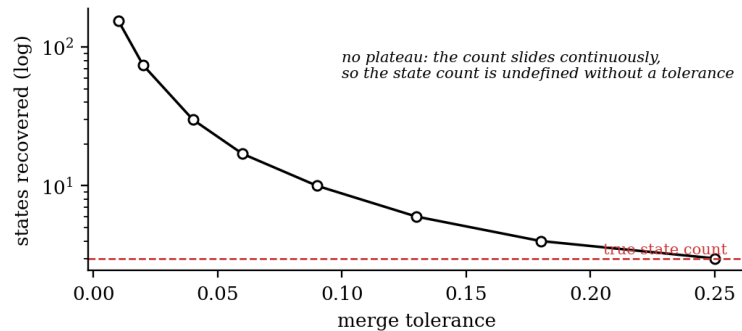


Figure 4. Recovered state count against merge tolerance. A system with genuinely discrete structure would show a plateau at the true count. There is none.

We read this as a finding rather than a tuning failure. A trained network's causal states are **approximate, not exact**: the interventional signatures form a continuum with a few dominant modes, which is why supplying the count recovers the structure almost perfectly while inferring it does not. The consequence is that a claim of the form "this model has N features" is undefined without a stated tolerance, and that the honest output of such an analysis is a rate-distortion curve -- description size against fidelity -- rather than a single number.

7. Scope and relation to existing work

This is a controlled demonstration, not a claim of novelty, and the distinction matters enough to state plainly. Interventional methods in interpretability are established, and the ideas surrounding this

measurement overlap existing literature: causal abstraction and interchange interventions provide the interventional framing, and computational mechanics supplies causal states as a minimal sufficient statistic, including recent work identifying belief-state geometry inside trained transformers.

What is offered here is narrower: a clean, reproducible measurement of how far the two quantities diverge on a system where the correct answer is known beforehand, together with an explicit negative result about state-count selection that we have not seen stated directly.

The practical implication is a methodological one. A probe result establishes presence. Only an intervention establishes use. Reporting the first as though it were the second is a category error, and on a system engineered to separate them it is wrong by a factor of roughly 1600.

8. Reproducibility

NumPy only, no machine-learning frameworks. Logistic regression, mutual information, adjusted Rand index and expected calibration error are implemented from scratch, so no library convention is assumed. Five seeds; all reported figures are means with standard deviations.

References

1. C. R. Shalizi and J. P. Crutchfield. Computational mechanics: pattern and prediction, structure and simplicity. *Journal of Statistical Physics*, 104:817-879, 2001.
2. A. Geiger, H. Lu, T. Icard and C. Potts. Causal abstractions of neural networks. *NeurIPS*, 2021.
3. A. S. Shai, S. E. Marzen, L. Teixeira, A. G. Oldenziel and P. M. Riechers. Transformers represent belief state geometry in their residual stream. *NeurIPS*, 2024.
4. L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2:193-218, 1985.