Causal Emergence: When Distortions in the Map Obscure the Territory

Frederick Eberhardt & Lin Lin Lee
Causal Representation Learning

- Equatorial Pacific Zonal Wind
- Equatorial Pacific SST

**westerly equatorial winds** → **causal relations** → **El Nino**

- Longitude: 160° to 210° to 260°
- Latitude: -8° to 8°
Mental Causation: Psychology vs. Neuroscience

mental state → brain state → mental state → brain state
Mental Causation: Psychology vs. Neuroscience

constitutive relation

mental state

brain state

causal relation

mental state

brain state
Micro- and Macro Causal Description

\[ f(I) = C \]

\[ P(\text{do}(I)) \]

\[ P(\text{do}(C)) \]

\[ P(E|\text{do}(C)) \]

\[ P(J|\text{do}(I)) \]

\[ T = g(J) \]
Hoel (2017): When the Map is Better than the Territory

finite state space

same statespace as I
Hoel (2017): When the Map is Better than the Territory

\[
\begin{bmatrix}
\frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & 0 & 0 & 0 \\
\frac{1}{7} & \frac{3}{7} & \frac{1}{7} & 0 & \frac{1}{7} & 0 & \frac{1}{7} & 0 \\
0 & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & 0 \\
\frac{1}{7} & 0 & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & 2/7 & 0 \\
1/9 & 2/9 & 2/9 & 1/9 & 0 & 2/9 & 1/9 & 0 \\
1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 \\
1/6 & 1/6 & 0 & 1/6 & 1/6 & 1/6 & 1/6 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}
\]

[Transition Probability Matrix]

finite state space

same statespace as I
Hoel (2017): *When the Map is Better than the Territory*

The finite state space of I is connected to the same state space of E through a transition probability matrix. The goal is to maximize the effective information:

$$\begin{bmatrix}
\frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & 0 & 0 & 0 \\
\frac{1}{7} & \frac{3}{7} & \frac{1}{7} & 0 & \frac{1}{7} & 0 & 1/7 & 0 \\
0 & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & 1/6 & 0 \\
\frac{1}{7} & 0 & 1/7 & \frac{1}{7} & \frac{1}{7} & 1/7 & 2/7 & 0 \\
\frac{1/9} & \frac{2/9} & \frac{2/9} & \frac{1/9} & 0 & \frac{2/9} & 1/9 & 0 \\
\frac{1/7} & \frac{1/7} & \frac{1/7} & \frac{1/7} & \frac{1/7} & \frac{1/7} & \frac{1/7} & 0 \\
\frac{1/6} & \frac{1/6} & 0 & \frac{1/6} & \frac{1/6} & \frac{1/6} & \frac{1/6} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}$$

[Transition Probability Matrix]
Effective Information $E_1$
Effective Information $EI$

\[ P(do(I)) = MaxEnt(I) \]

\[ E(J) = \frac{1}{n} \sum_{I} P(J|do(I)) \]

intervention distribution

[Transition Probability Matrix]

finite state space

same state space as $I$
Effective Information $EI$

Difference between effect of specific intervention and (maxent) average intervention:

\[ P(J|do(I = i)) \] vs. \[ E(J) \]

intervention distribution

\[ P(do(I)) = MaxEnt(I) \]

effect distribution

\[ E(J) = \frac{1}{n} \sum_I P(J|do(I)) \]

[Transition Probability Matrix]

finite state space

same state space as $I$
Effective Information $EI$

$$EI(I \rightarrow J) = \sum_{I} P(do(I)) D_{KL}(P(J|do(I)) || E(J))$$

KL-divergence

Difference between effect of specific intervention and (maxent) average intervention:

$$P(J|do(I = i)) \quad \text{vs.} \quad E(J)$$

intervention distribution
$$P(do(I)) = \text{MaxEnt}(I)$$

effect distribution
$$E(J) = \frac{1}{n} \sum_{I} P(J|do(I))$$

*this slide has been corrected for a typo that was in the original*
Effective Information $EI$:

$$EI(I \rightarrow J) = I(I_{maxEnt}, J_E)$$

$$= \sum_I P(do(I))D_{KL}(P(J|do(I))\|E(J))$$

KL-divergence

Difference between effect of specific intervention and (maxent) average intervention:

$$P(J|do(I = i)) \text{ vs. } E(J)$$

intervention distribution

$$P(do(I)) = MaxEnt(I)$$

effect distribution

$$E(J) = \frac{1}{n} \sum_I P(J|do(I))$$

*this slide has been corrected for a typo that was in the original*
What is great about Effective Information?

\[ EI(I \to J) = I(I_{\text{maxEnt}}, J_E) \]

\[ = \sum_I P(\text{do}(I)) D_{KL}(P(J|\text{do}(I)) \mid \mid E(J)) \]

- **directed** information measure (defined in terms of interventions)
- connection between causality and information theory
- explores full cause space / is independent of observed \( P(I) \)
- [core feature of characterization of consciousness in Tononi’s Integrated Information Theory of Consciousness]

\[ \text{[Transition Probability Matrix]} \]

finite state space  \hspace{1cm} \text{same state space as } I

*this slide has been corrected for a typo that was in the original*
Hoel’s Causal Emergence

Finite state space

Transition Probability Matrix

$max_{C(I)} E I(I \rightarrow J)$

Coarsenings of state space of I

same state space as I

finite state space
Hoel’s Macro Intervention

$do(C = c)$

$max_{C(I)} EI(I \rightarrow J)$

coarsenings of state space of $I$

Transition Probability Matrix

same state space as $I$
Hoel’s Macro Intervention

\[ \text{do}(C = c) \]

\[ \max_{C(I)} EI(I \rightarrow J) \]

coarsenings of state space of \( I \)

same state space as \( I \)

\[ \frac{1}{n_i} \sum \text{do}(I) \]
Causal Emergence in Hoel 2017

\[ P(\text{do}(C)) = \text{MaxEnt}(C) \]

\[ P(E|\text{do}(C)) = \max_{C(\mathcal{I})} EI(I \rightarrow J) \]

\[ P(\text{do}(I)) = \frac{1}{\eta} \sum_{i} \text{do}(i) \]

\[ P(J|\text{do}(I)) \]
Causal Emergence in Hoel 2017

\[ P(\text{do}(C)) = \text{MaxEnt}(C) \]

\[ P(E|\text{do}(C)) \]

\[ \max_{C(I)} EI(I \rightarrow J) \]

\[ \approx \max_{P(I)} I(I, J) = Ch(I, J) \]

\[ P(\text{do}(I)) \]

\[ P(J|\text{do}(I)) \]

channel capacity

mutual information
Causal Emergence in Hoel 2017

\[ P(\text{do}(C)) = \text{MaxEnt}(C) \]

\[ P(E|\text{do}(C)) \]

\[ \max_{C(I \rightarrow J)} EI(I \rightarrow J) \]

\[ \approx \max_{P(I)} I(I, J) = Ch(I, J) \]

\[ P(\text{do}(I)) \]

\[ P(J|\text{do}(I)) \]
Causal Emergence in Hoel 2017

\[
P(\text{do}(C)) = \text{MaxEnt}(C)
\]

\[
P(E|\text{do}(C))
\]

**BUT**, the EI-maximization is subject to:

- subset of possible intervention distributions
- identical state spaces for I and J that change with the coarsening

\[
\max_{C(\mathcal{I})} EI(I \rightarrow J)
\]

\[
\approx \max_{P(I)} I(I, J) = Ch(I, J)
\]

\[
P(\text{do}(I))
\]

\[
P(J|\text{do}(I))
\]

causal capacity

mutual information

channel capacity
Example 1

$$\max_{C(I)} EI(I \rightarrow J)$$

$$\begin{bmatrix}
1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 \\
1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 \\
1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 \\
1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 \\
1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 \\
1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 \\
1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 \\
1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}$$
Example 1

\[ \max_{C(I)} EI(I \rightarrow J) \]

\[
\begin{bmatrix}
1 & 0 \\
0 & 1 \\
\end{bmatrix}
\]

\[
\begin{bmatrix}
1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 \\
1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 \\
1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 \\
1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 \\
1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 \\
1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 \\
1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 \\
1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}
\]
Example 2

$$\max_{C(I)} EI(I \rightarrow J)$$

$$\begin{bmatrix}
1/5 & 1/5 & 1/5 & 1/5 & 1/5 & 0 & 0 & 0 \\
1/7 & 3/7 & 1/7 & 0 & 1/7 & 0 & 1/7 & 0 \\
0 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 0 \\
1/7 & 0 & 1/7 & 1/7 & 1/7 & 1/7 & 2/7 & 0 \\
1/9 & 2/9 & 2/9 & 1/9 & 0 & 2/9 & 1/9 & 0 \\
1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 \\
1/6 & 1/6 & 0 & 1/6 & 1/6 & 1/6 & 1/6 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}$$
Example 2

$$\max_{C(I)} EI(I \rightarrow J)$$

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 1/5 & 1/5 & 1/5 & 1/5 & 1/5 & 0 & 0 & 0 \\
1/7 & 3/7 & 1/7 & 0 & 1/7 & 0 & 1/7 & 0 \\
0 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 0 \\
1/7 & 0 & 1/7 & 1/7 & 1/7 & 1/7 & 2/7 & 0 \\
1/9 & 2/9 & 2/9 & 1/9 & 0 & 2/9 & 1/9 & 0 \\
1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 \\
1/6 & 1/6 & 0 & 1/6 & 1/6 & 1/6 & 1/6 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$
Example 2

\[
\begin{bmatrix}
1 & 0 \\
0 & 1
\end{bmatrix}
\]

\[
\max_{C(I)} EI(I \to J)
\]

\[
\begin{bmatrix}
1/5 & 1/5 & 1/5 & 1/5 & 1/5 & 0 & 0 & 0 \\
1/7 & 3/7 & 1/7 & 0 & 1/7 & 0 & 1/7 & 0 \\
0 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 0 \\
1/7 & 0 & 1/7 & 1/7 & 1/7 & 1/7 & 2/7 & 0 \\
1/9 & 2/9 & 2/9 & 1/9 & 1/9 & 0 & 2/9 & 1/9 \\
1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 \\
1/6 & 1/6 & 0 & 1/6 & 1/6 & 1/6 & 1/6 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}
\]
Ambiguous Manipulation

Total Cholesterol $\rightarrow$ Heart Disease

Spirtes & Scheines (2002)
Ambiguous Manipulation

Total Cholesterol

HDLC

LDL

+ 

-

Heart Disease

Spirtes & Scheines (2002)
the causal effect of *Total Cholesterol* on *Heart Disease* is **ambiguous**

⇒ *Total Cholesterol* is over-aggregated, it cannot be described as a cause of *Heart Disease*
Example 2

How different can the causal effects of two micro states be such that they still get mapped to the same macro state?
Example 3: collapsing micro states with different causal effects

\[
\max_{C(I)} EI(I \rightarrow J)
\]

\[
\begin{bmatrix}
0.85 & 0.15 \\
0 & 1
\end{bmatrix}
\]

\[
\begin{bmatrix}
0.7 & 0.1 & 0.2 \\
0.4 & 0.5 & 0.1 \\
0 & 0 & 1
\end{bmatrix}
\]
Example 3: collapsing micro states with different causal effects

\[
\begin{bmatrix}
0.85 & 0.15 \\
0 & 1
\end{bmatrix}
\]

\[
\begin{bmatrix}
0.7 & 0.1 & 0.2 \\
0.4 & 0.5 & 0.1 \\
0 & 0 & 1
\end{bmatrix}
\]

- worst case examples obviously depend on distance metric for distinct states.
- importantly, if the first entry had been 0.8, then the first two states would not have been collapsed.

What is the difference between mixtures and macro variables?
Marginalization
Marginalization
Marginalization
Abstraction and Marginalization should **commute**

\[ C \rightarrow C' \rightarrow E \]

\[ I \rightarrow I' \rightarrow J \]

**abstraction**

**marginalization**

**abstraction**

**marginalization**
Abstraction and Marginalization DO NOT commute in Hoel 2017

\[ \begin{bmatrix} 0.5 & 0.5 \\ 1 & 0 \end{bmatrix} \quad \begin{bmatrix} 0.5 & 0.5 \\ 1 & 0 \end{bmatrix} \]

\[ \begin{bmatrix} 0.3 & 0.2 & 0.5 \\ 0.4 & 0.1 & 0.5 \\ 0.5 & 0.5 & 0 \end{bmatrix} \quad \begin{bmatrix} 0.3 & 0.2 & 0.5 \\ 0.4 & 0.1 & 0.5 \\ 0.5 & 0.5 & 0 \end{bmatrix} \]
Abstraction and Marginalization DO NOT commute in Hoel 2017

\[
\begin{bmatrix}
0.75 & 0.25 \\
0.5 & 0.5
\end{bmatrix}
\]

\[
\begin{bmatrix}
0.3 & 0.2 & 0.5 \\
0.4 & 0.1 & 0.5 \\
0.5 & 0.5 & 0
\end{bmatrix}
\]

\[
\begin{bmatrix}
0.3 & 0.2 & 0.5 \\
0.4 & 0.1 & 0.5 \\
0.5 & 0.5 & 0
\end{bmatrix}
\]
Abstraction and Marginalization DO NOT commute in Hoel 2017
Abstraction and Marginalization DO NOT commute in Hoel 2017

\[
\begin{bmatrix}
0.42 & 0.33 & 0.25 \\
0.41 & 0.34 & 0.25 \\
0.35 & 0.15 & 0.5
\end{bmatrix}
\]
Abstraction and Marginalization DO NOT commute in Hoel 2017

effective information is already maximized at the micro level
Abstraction and Marginalization DO NOT commute in Hoel 2017

\[
\begin{bmatrix}
0.75 & 0.25 \\
0.5 & 0.5 \\
\end{bmatrix}
\]

\[\neq\]

\[
\begin{bmatrix}
0.42 & 0.33 & 0.25 \\
0.41 & 0.34 & 0.25 \\
0.35 & 0.15 & 0.5 \\
\end{bmatrix}
\]

effective information is already maximized at the micro level
The Problem: Introducing MaxEnt distributions

\[ P(\text{do}(I')) = \text{MaxEnt}(I') \]
\[ P(\text{do}(I)) = \text{MaxEnt}(I) \]

\[ P(\text{do}(I)) = \text{MaxEnt}(I) \neq P(I' | \text{do}(I)) \]
Upshots

- it is worth distinguishing between macro level causes (or causal representations) and mixtures of causal effects
- whether or not there are macro-level causal descriptions is an empirical question determined by $P(E \mid do(C))$, independent of $P(do(C))$
Upshots

- it is worth distinguishing between **macro level causes** (or causal representations) and **mixtures of causal effects**

- whether or not there are macro-level causal descriptions is an empirical question determined by $P(E | do(C))$, independent of $P(do(C))$

  ➡ this also ensures that abstraction and marginalization commute
Upshots

• it is worth distinguishing between macro level causes (or causal representations) and mixtures of causal effects

• whether or not there are macro-level causal descriptions is an empirical question determined by $P(E | \text{do}(C))$, independent of $P(\text{do}(C))$

→ this also ensures that abstraction and marginalization commute

• (although I have not discussed this in detail here) there is a distinction between how one determines the macro cause and how one determines the macro effect, though of course they are related
Specifically for Hoel’s account

- the suggested relation between information theory and causality via effective information is tenuous and suggestive at best
Specifically for Hoel’s account

- the suggested relation between information theory and causality via effective information is tenuous and suggestive at best

- channel capacity is a normative concept; whether or not it is exhausted is an empirical question; so the described causal emergence here is a possible emergence that may never be exhibited by the system in question

- effective information is uniquely maximized, but it is not clear that the implied partition of the state space is unique; this cuts both ways: either one wants uniqueness, or one wants non-uniqueness but not in the way implied by this theory: one wants many very different levels of aggregation
References

- Erik P Hoel. When the map is better than the territory. Entropy, 19(5):188, 2017.

Other useful references
- Paul Rubenstein, Sebastian Weichwald et al. Causal consistency of structural equation models. UAI 2017
- Scott Aaronson, Higher-level causation exists (but I wish it didn’t). https://www.scottaaronson.com/blog/?p=3294 (and reply by Hoel)

Attempts at an alternative account (i.e. shameless self-promotion)

Thank you!