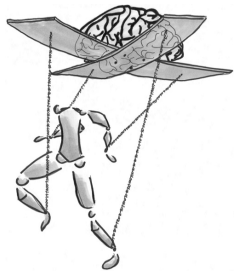
Causal Interpretation Rules for Encoding and Decoding Models in Neuroimaging

Moritz Grosse-Wentrup

Max Planck Institute for Intelligent Systems
Department Empirical Inference
Tübingen, Germany

June 14, 2015



Causal Terminology in Neuroimaging

Causal Terminology in Neuroimaging

Hippocampal activity in this study was **correlated** *with amygdala activity, supporting the view that the amygdala* **enhances** *explicit memory by* **modulating** *activity in the hippocampus.*

(Anonymous author, *Trends in Cognitive Sciences*, 2001)

Causal Terminology in Neuroimaging

Hippocampal activity in this study was **correlated** *with amygdala activity, supporting the view that the amygdala* **enhances** *explicit memory by* **modulating** *activity in the hippocampus.*

(Anonymous author, *Trends in Cognitive Sciences*, 2001)

We tested [...] whether pre-stimulus alpha oscillations **measured** *with electroencephalography (EEG)* **influence** *the encoding of items into working memory.*

(Anonymous authors, *Journal of Neuroscience*, 2014)

Causal Terminology in Neuroimaging

Hippocampal activity in this study was **correlated** *with amygdala activity, supporting the view that the amygdala* **enhances** *explicit memory by* **modulating** *activity in the hippocampus.*

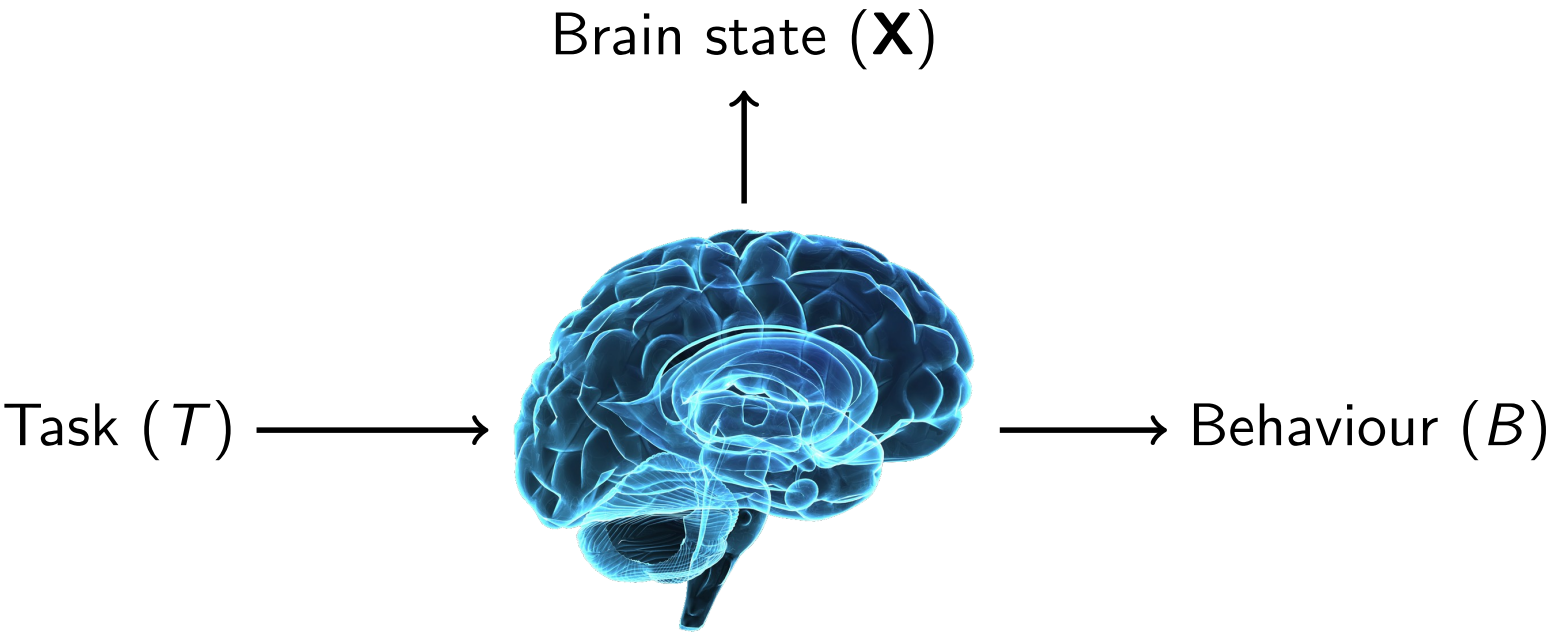
(Anonymous author, *Trends in Cognitive Sciences*, 2001)

We tested [...] whether pre-stimulus alpha oscillations **measured** *with electroencephalography (EEG)* **influence** *the encoding of items into working memory.*

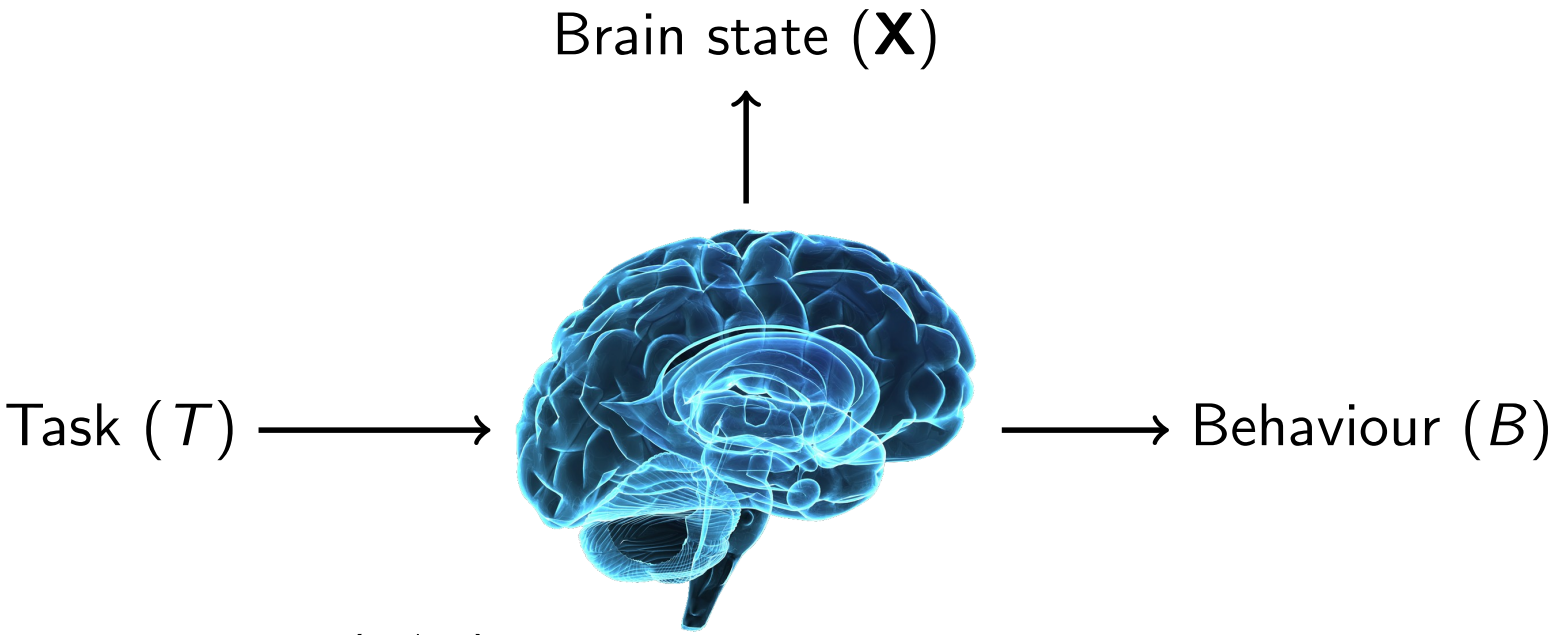
(Anonymous authors, *Journal of Neuroscience*, 2014)

→ Which causal statements are warranted and which ones are not supported by empirical evidence?

Causal Modelling in Neuroimaging

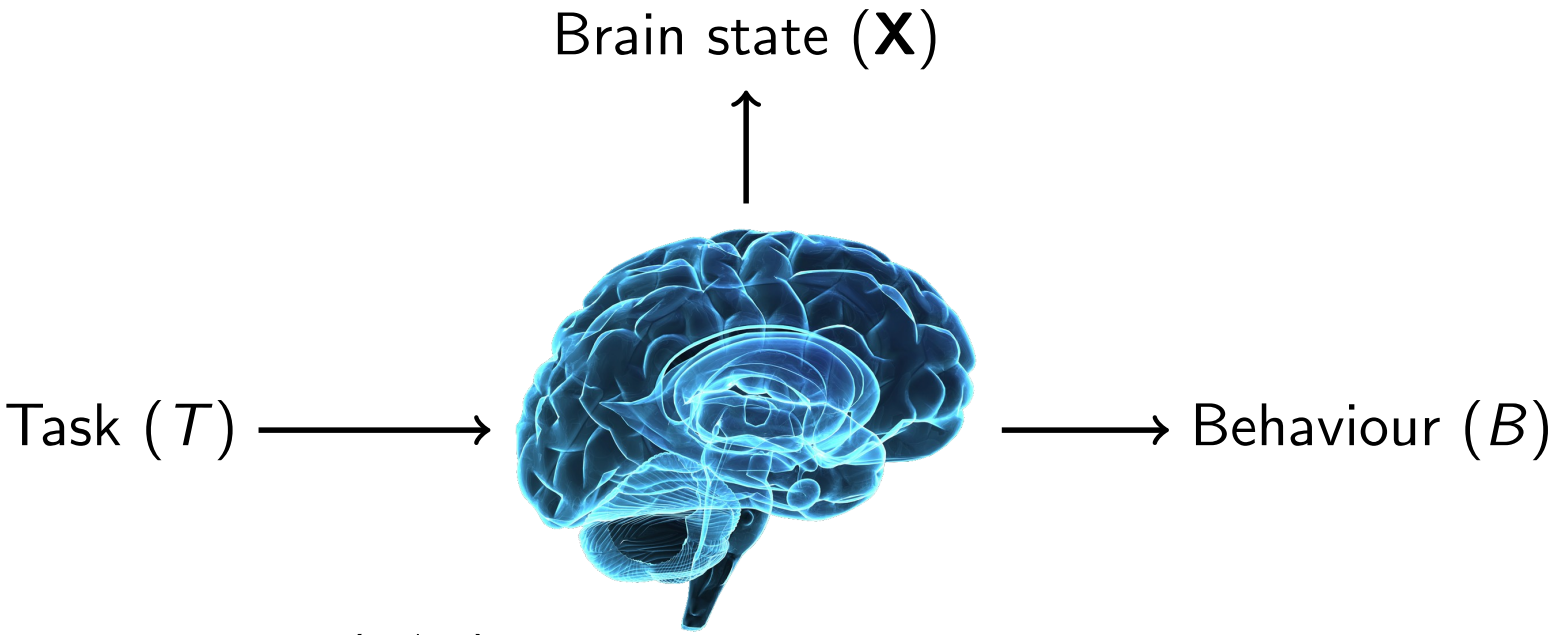


Causal Modelling in Neuroimaging



Encoding models: $p(\mathbf{X}|C)$

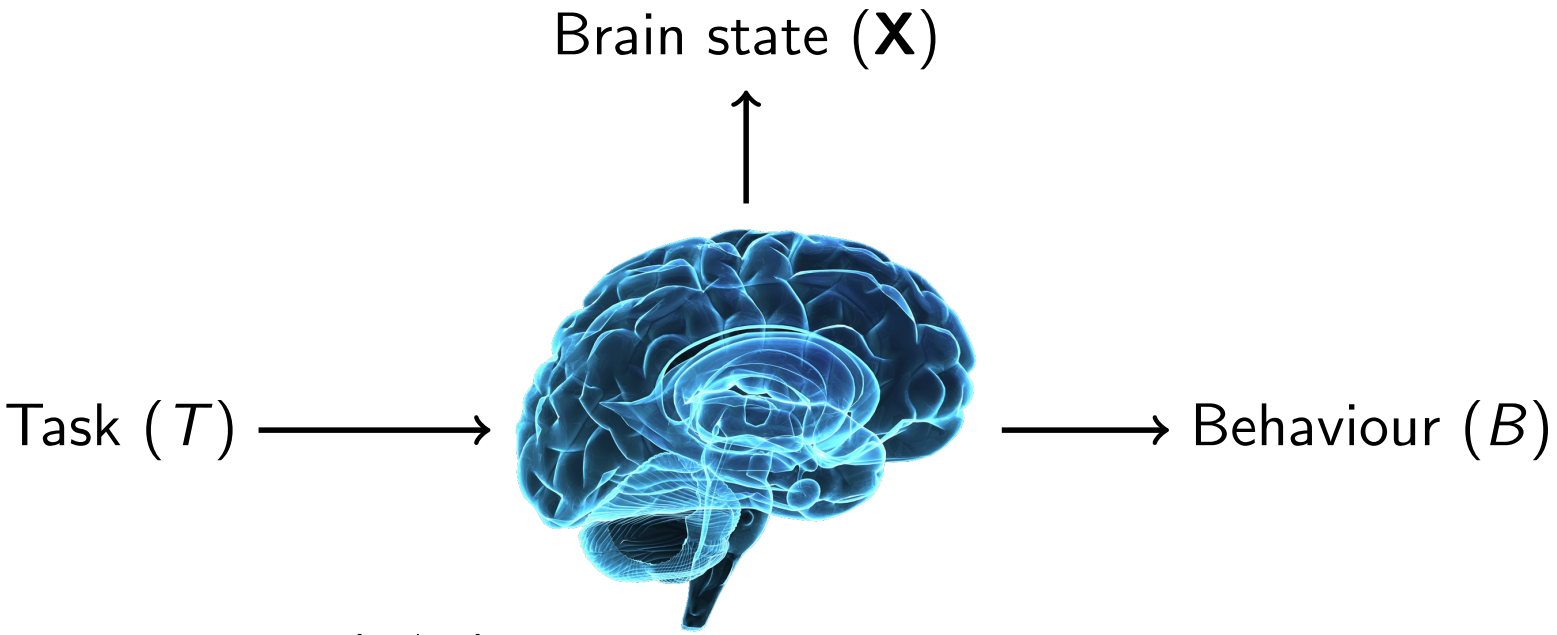
Causal Modelling in Neuroimaging



Encoding models: $p(\mathbf{X}|C)$

- Task-based: $p(\mathbf{X}|T) \equiv p(\text{effect}|\text{cause}) \Rightarrow$ causal direction

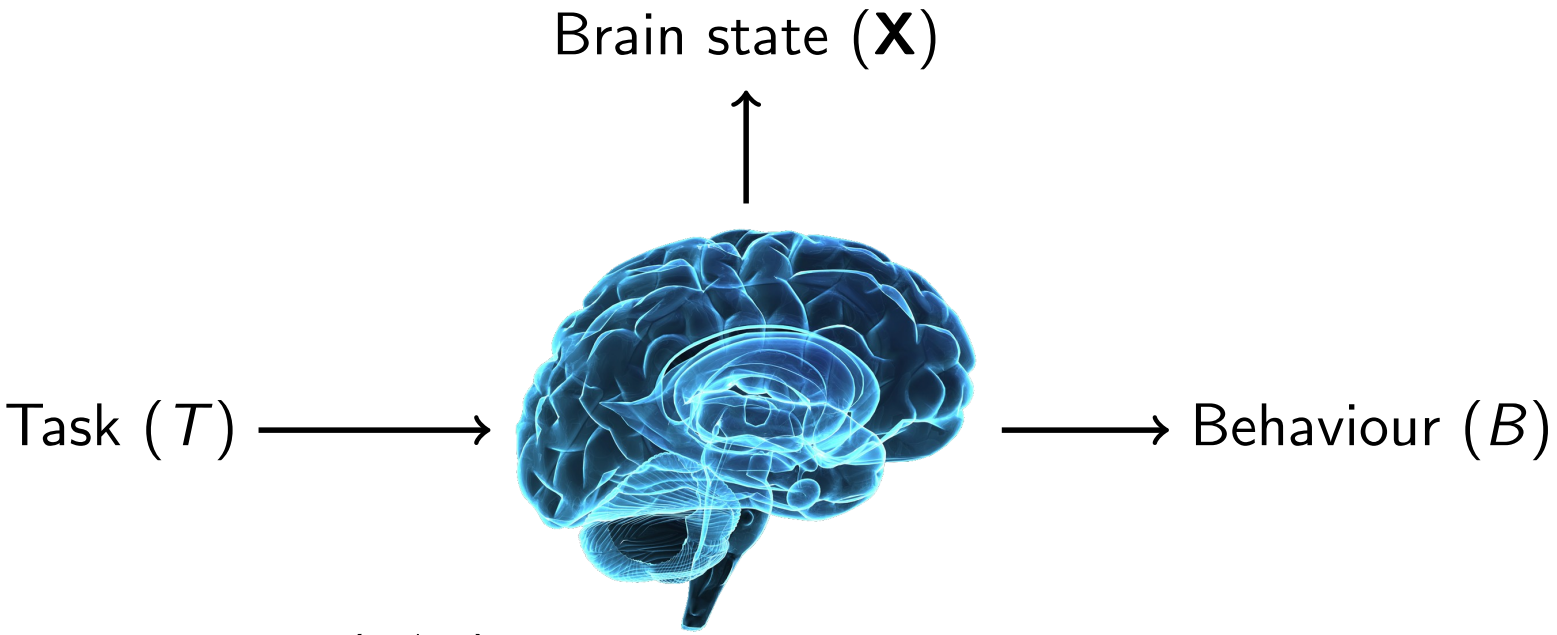
Causal Modelling in Neuroimaging



Encoding models: $p(\mathbf{X}|C)$

- Task-based: $p(\mathbf{X}|T) \equiv p(\text{effect}|\text{cause}) \Rightarrow$ causal direction
- Behaviour-based: $p(\mathbf{X}|B) \equiv p(\text{cause}|\text{effect}) \Rightarrow$ anti-causal direction

Causal Modelling in Neuroimaging

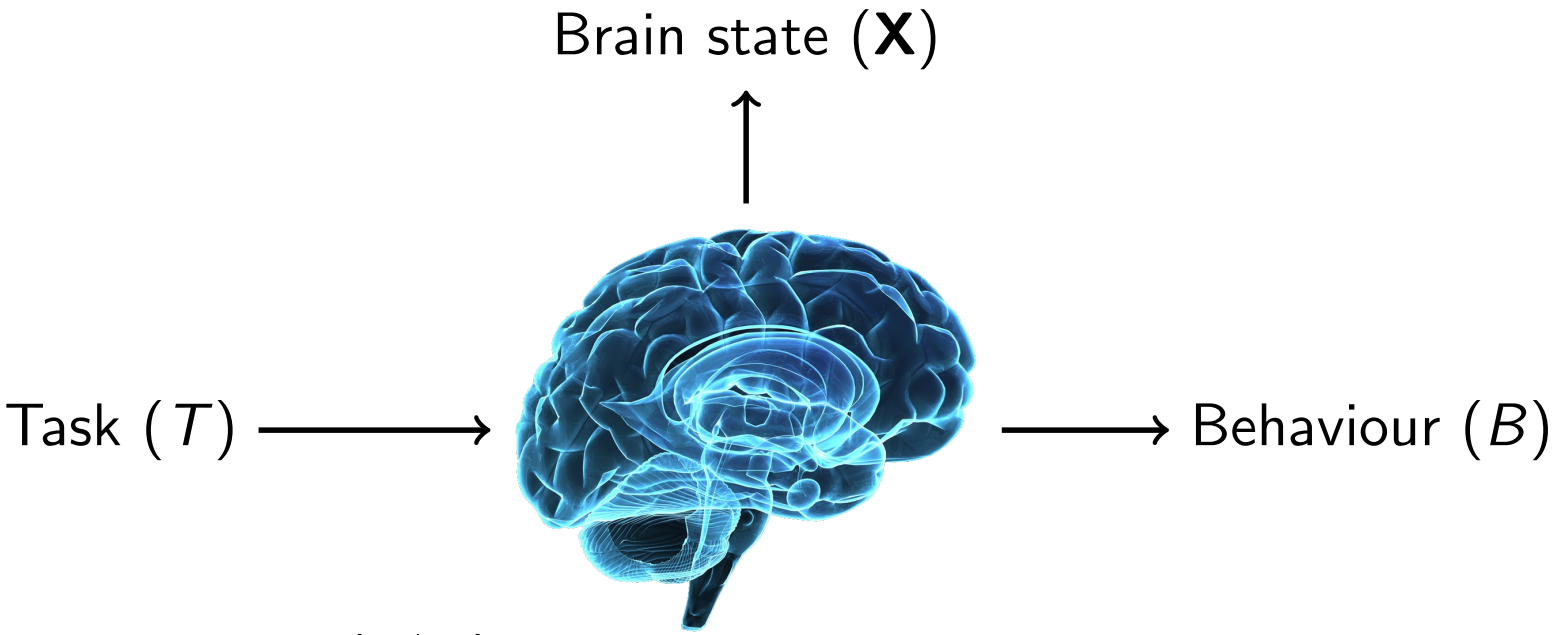


Encoding models: $p(\mathbf{X}|C)$

- Task-based: $p(\mathbf{X}|T) \equiv p(\text{effect}|\text{cause}) \Rightarrow$ causal direction
- Behaviour-based: $p(\mathbf{X}|B) \equiv p(\text{cause}|\text{effect}) \Rightarrow$ anti-causal direction

Decoding models: $p(C|\mathbf{X})$

Causal Modelling in Neuroimaging



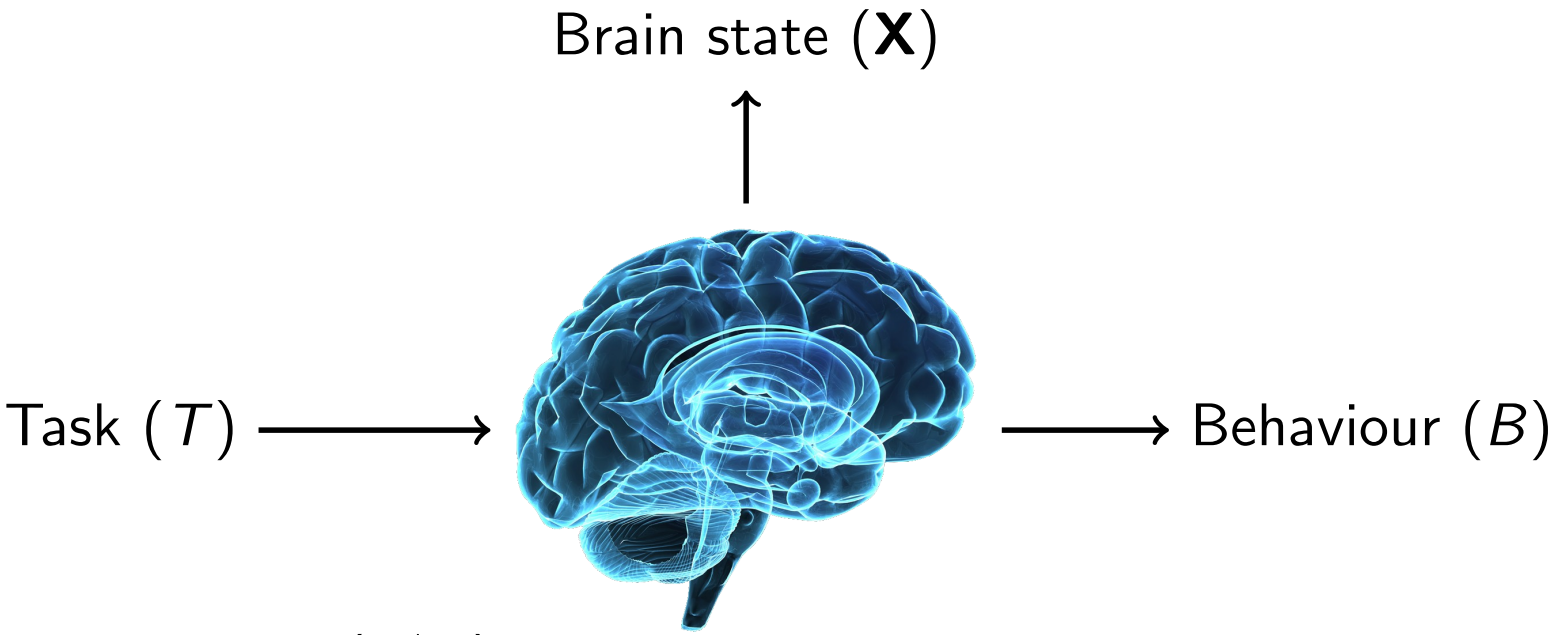
Encoding models: $p(\mathbf{X}|C)$

- Task-based: $p(\mathbf{X}|T) \equiv p(\text{effect}|\text{cause}) \Rightarrow$ causal direction
- Behaviour-based: $p(\mathbf{X}|B) \equiv p(\text{cause}|\text{effect}) \Rightarrow$ anti-causal direction

Decoding models: $p(C|\mathbf{X})$

- Task-based: $p(T|\mathbf{X}) \equiv p(\text{cause}|\text{effect}) \Rightarrow$ anti-causal direction

Causal Modelling in Neuroimaging



Encoding models: $p(\mathbf{X}|C)$

- Task-based: $p(\mathbf{X}|T) \equiv p(\text{effect}|\text{cause}) \Rightarrow$ causal direction
- Behaviour-based: $p(\mathbf{X}|B) \equiv p(\text{cause}|\text{effect}) \Rightarrow$ anti-causal direction

Decoding models: $p(C|\mathbf{X})$

- Task-based: $p(T|\mathbf{X}) \equiv p(\text{cause}|\text{effect}) \Rightarrow$ anti-causal direction
- Behaviour-based: $p(B|\mathbf{X}) \equiv p(\text{effect}|\text{cause}) \Rightarrow$ causal direction

Notation

Notation

- Experimental condition C

Notation

- Experimental condition C, such as
  - type of stimulus displayed to a subject: $C = T \in \{-1, +1\}$

Notation

- Experimental condition C, such as
  - type of stimulus displayed to a subject: $C = T \in \{-1, +1\}$
  - delay of a behavioural response: $C = B \in \mathbb{R}_+$

Notation

- Experimental condition C, such as
  - type of stimulus displayed to a subject: $C = T \in \{-1, +1\}$
  - delay of a behavioural response: $C = B \in \mathbb{R}_+$
  - ...

Notation

- Experimental condition C, such as
  - type of stimulus displayed to a subject: $C = T \in \{-1, +1\}$
  - delay of a behavioural response: $C = B \in \mathbb{R}_+$
  - ...
- Brain state features $\mathbf{X} \in \mathbb{R}^M$, such as

Notation

- Experimental condition C, such as
  - type of stimulus displayed to a subject: $C = T \in \{-1, +1\}$
  - delay of a behavioural response: $C = B \in \mathbb{R}_+$
  - ...
- Brain state features $\mathbf{X} \in \mathbb{R}^M$, such as
  - BOLD response at M voxels measured by fMRI

Notation

- Experimental condition C, such as
  - type of stimulus displayed to a subject: $C = T \in \{-1, +1\}$
  - delay of a behavioural response: $C = B \in \mathbb{R}_+$
  - ...
- Brain state features $\mathbf{X} \in \mathbb{R}^M$, such as
  - BOLD response at M voxels measured by fMRI
  - bandpower at M EEG channels

Notation

- Experimental condition C, such as
  - type of stimulus displayed to a subject: $C = T \in \{-1, +1\}$
  - delay of a behavioural response: $C = B \in \mathbb{R}_+$
  - ...
- Brain state features $\mathbf{X} \in \mathbb{R}^M$, such as
  - BOLD response at M voxels measured by fMRI
  - bandpower at M EEG channels
  - mean spike-rate of M neurons

Notation

- Experimental condition C, such as
  - type of stimulus displayed to a subject: $C = T \in \{-1, +1\}$
  - delay of a behavioural response: $C = B \in \mathbb{R}_+$
  - ...
- Brain state features $\mathbf{X} \in \mathbb{R}^M$, such as
  - BOLD response at M voxels measured by fMRI
  - bandpower at M EEG channels
  - mean spike-rate of M neurons
  - ...

Notation

- Experimental condition C, such as
  - type of stimulus displayed to a subject: $C = T \in \{-1, +1\}$
  - delay of a behavioural response: $C = B \in \mathbb{R}_+$
  - ...
- Brain state features $\mathbf{X} \in \mathbb{R}^M$, such as
  - BOLD response at M voxels measured by fMRI
  - bandpower at M EEG channels
  - mean spike-rate of M neurons
  - ...
- Experimental data $\{(c, \mathbf{x})_1, \ldots, (c, \mathbf{x})_N\}$, drawn i.i.d. from $p(C, \mathbf{X})$

Notation

- Experimental condition C, such as
  - type of stimulus displayed to a subject: $C = T \in \{-1, +1\}$
  - delay of a behavioural response: $C = B \in \mathbb{R}_+$
  - ...
- Brain state features $\mathbf{X} \in \mathbb{R}^M$, such as
  - BOLD response at M voxels measured by fMRI
  - bandpower at M EEG channels
  - mean spike-rate of M neurons
  - ...
- Experimental data $\{(c, \mathbf{x})_1, \ldots, (c, \mathbf{x})_N\}$, drawn i.i.d. from $p(C, \mathbf{X})$
- Assumption: Oracle for properties of $p(C, \mathbf{X})$ available

Notation

- Experimental condition C, such as
  - type of stimulus displayed to a subject: $C = T \in \{-1, +1\}$
  - delay of a behavioural response: $C = B \in \mathbb{R}_+$
  - ...
- Brain state features $\mathbf{X} \in \mathbb{R}^M$, such as
  - BOLD response at M voxels measured by fMRI
  - bandpower at M EEG channels
  - mean spike-rate of M neurons
  - ...
- Experimental data $\{(c, \mathbf{x})_1, \ldots, (c, \mathbf{x})_N\}$, drawn i.i.d. from $p(C, \mathbf{X})$
- Assumption: Oracle for properties of $p(C, \mathbf{X})$ available
  - Statistical independence:
    $C \perp\!\!\!\perp X_i \Leftrightarrow p(C, X_i) = p(C)p(X_i)$

Notation

- Experimental condition C, such as
  - type of stimulus displayed to a subject: $C = T \in \{-1, +1\}$
  - delay of a behavioural response: $C = B \in \mathbb{R}_+$
  - ...
- Brain state features $\mathbf{X} \in \mathbb{R}^M$, such as
  - BOLD response at M voxels measured by fMRI
  - bandpower at M EEG channels
  - mean spike-rate of M neurons
  - ...
- Experimental data $\{(c, \mathbf{x})_1, \ldots, (c, \mathbf{x})_N\}$, drawn i.i.d. from $p(C, \mathbf{X})$
- Assumption: Oracle for properties of $p(C, \mathbf{X})$ available
  - Statistical independence:
    $C \perp\!\!\!\perp X_i \Leftrightarrow p(C, X_i) = p(C)p(X_i)$
  - Conditional statistical independence:
    $C \perp\!\!\!\perp X_i | \mathbf{X} \backslash X_i \Leftrightarrow p(C, X_i | \mathbf{X} \backslash X_i) = p(C | \mathbf{X} \backslash X_i) p(X_i | \mathbf{X} \backslash X_i)$

Causal Bayesian Networks (Pearl, 2000; Spirtes et al., 2000)

Causal Bayesian Networks (Pearl, 2000; Spirtes et al., 2000)

X_i is a cause of X_j ($X_i \rightarrow X_j$), iff there exist values of X_i and X_j such that $p(x_j|\text{do}\{x_i\}) \neq p(x_j)$.

Causal Bayesian Networks (Pearl, 2000; Spirtes et al., 2000)

X_i is a cause of X_j ($X_i \rightarrow X_j$), iff there exist values of X_i and X_j such that $p(x_j|\text{do}\{x_i\}) \neq p(x_j)$.

The chain

$$X_1 \rightarrow X_2 \rightarrow X_3$$

Causal Bayesian Networks (Pearl, 2000; Spirtes et al., 2000)

X_i is a cause of X_j ($X_i \rightarrow X_j$), iff there exist values of X_i and X_j such that $p(x_j|\text{do}\{x_i\}) \neq p(x_j)$.

The chain	The fork
$X_1 \rightarrow X_2 \rightarrow X_3$	$X_1 \leftarrow X_2 \rightarrow X_3$

Causal Bayesian Networks (Pearl, 2000; Spirtes et al., 2000)

X_i is a cause of X_j ($X_i \rightarrow X_j$), iff there exist values of X_i and X_j such that $p(x_j|\text{do}\{x_i\}) \neq p(x_j)$.

The chain	The fork	The collider
$X_1 \rightarrow X_2 \rightarrow X_3$	$X_1 \leftarrow X_2 \rightarrow X_3$	$X_1 \rightarrow X_2 \leftarrow X_3$

Causal Bayesian Networks (Pearl, 2000; Spirtes et al., 2000)

X_i is a cause of X_j ($X_i \to X_j$), iff there exist values of X_i and X_j such that $p(x_j|\text{do}\{x_i\}) \neq p(x_j)$.

The chain	The fork	The collider
$X_1 \to X_2 \to X_3$	$X_1 \leftarrow X_2 \to X_3$	$X_1 \to X_2 \leftarrow X_3$

- Causal Markov condition: Independence relations implied by a directed acyclic graph (DAG) are encoded in every $p(\mathbf{X})$ generated by this DAG.

Causal Bayesian Networks (Pearl, 2000; Spirtes et al., 2000)

X_i is a cause of X_j ($X_i \rightarrow X_j$), iff there exist values of X_i and X_j such that $p(x_j|\text{do}\{x_i\}) \neq p(x_j)$.

The chain	The fork	The collider
$X_1 \rightarrow X_2 \rightarrow X_3$	$X_1 \leftarrow X_2 \rightarrow X_3$	$X_1 \rightarrow X_2 \leftarrow X_3$
$X_1 \not\perp\!\!\!\perp X_3$		

- Causal Markov condition: Independence relations implied by a directed acyclic graph (DAG) are encoded in every $p(\mathbf{X})$ generated by this DAG.

Causal Bayesian Networks (Pearl, 2000; Spirtes et al., 2000)

X_i is a cause of X_j ($X_i \rightarrow X_j$), iff there exist values of X_i and X_j such that $p(x_j|\text{do}\{x_i\}) \neq p(x_j)$.

The chain	The fork	The collider
$X_1 \rightarrow X_2 \rightarrow X_3$	$X_1 \leftarrow X_2 \rightarrow X_3$	$X_1 \rightarrow X_2 \leftarrow X_3$
$X_1 \not\perp\!\!\!\perp X_3$		
$X_1 \perp\!\!\!\perp X_3 \mid X_2$		

- Causal Markov condition: Independence relations implied by a directed acyclic graph (DAG) are encoded in every $p(\mathbf{X})$ generated by this DAG.

Causal Bayesian Networks (Pearl, 2000; Spirtes et al., 2000)

X_i is a cause of X_j ($X_i \rightarrow X_j$), iff there exist values of X_i and X_j such that $p(x_j|\text{do}\{x_i\}) \neq p(x_j)$.

The chain	The fork	The collider
$X_1 \rightarrow X_2 \rightarrow X_3$	$X_1 \leftarrow X_2 \rightarrow X_3$	$X_1 \rightarrow X_2 \leftarrow X_3$
$X_1 \not\perp\!\!\!\perp X_3$	$X_1 \not\perp\!\!\!\perp X_3$	
$X_1 \perp\!\!\!\perp X_3 \mid X_2$		

- Causal Markov condition: Independence relations implied by a directed acyclic graph (DAG) are encoded in every $p(\mathbf{X})$ generated by this DAG.

Causal Bayesian Networks (Pearl, 2000; Spirtes et al., 2000)

X_i is a cause of X_j ($X_i \rightarrow X_j$), iff there exist values of X_i and X_j such that $p(x_j|\text{do}\{x_i\}) \neq p(x_j)$.

The chain	The fork	The collider
$X_1 \rightarrow X_2 \rightarrow X_3$	$X_1 \leftarrow X_2 \rightarrow X_3$	$X_1 \rightarrow X_2 \leftarrow X_3$
$X_1 \not\perp\!\!\!\perp X_3$	$X_1 \not\perp\!\!\!\perp X_3$	
$X_1 \perp\!\!\!\perp X_3 \mid X_2$	$X_1 \perp\!\!\!\perp X_3 \mid X_2$	

- Causal Markov condition: Independence relations implied by a directed acyclic graph (DAG) are encoded in every $p(\mathbf{X})$ generated by this DAG.

Causal Bayesian Networks (Pearl, 2000; Spirtes et al., 2000)

X_i is a cause of X_j ($X_i \to X_j$), iff there exist values of X_i and X_j such that $p(x_j|\text{do}\{x_i\}) \neq p(x_j)$.

The chain	The fork	The collider
$X_1 \to X_2 \to X_3$	$X_1 \leftarrow X_2 \to X_3$	$X_1 \to X_2 \leftarrow X_3$
$X_1 \not\perp\!\!\!\perp X_3$	$X_1 \not\perp\!\!\!\perp X_3$	$X_1 \perp\!\!\!\perp X_3$
$X_1 \perp\!\!\!\perp X_3 \mid X_2$	$X_1 \perp\!\!\!\perp X_3 \mid X_2$	

- Causal Markov condition: Independence relations implied by a directed acyclic graph (DAG) are encoded in every $p(\mathbf{X})$ generated by this DAG.

Causal Bayesian Networks (Pearl, 2000; Spirtes et al., 2000)

X_i is a cause of X_j ($X_i \rightarrow X_j$), iff there exist values of X_i and X_j such that $p(x_j|\text{do}\{x_i\}) \neq p(x_j)$.

The chain	The fork	The collider
$X_1 \rightarrow X_2 \rightarrow X_3$	$X_1 \leftarrow X_2 \rightarrow X_3$	$X_1 \rightarrow X_2 \leftarrow X_3$
$X_1 \not\perp\!\!\!\perp X_3$	$X_1 \not\perp\!\!\!\perp X_3$	$X_1 \perp\!\!\!\perp X_3$
$X_1 \perp\!\!\!\perp X_3 \mid X_2$	$X_1 \perp\!\!\!\perp X_3 \mid X_2$	$X_1 \not\perp\!\!\!\perp X_3 \mid X_2$

- Causal Markov condition: Independence relations implied by a directed acyclic graph (DAG) are encoded in every $p(\mathbf{X})$ generated by this DAG.

Causal Bayesian Networks (Pearl, 2000; Spirtes et al., 2000)

X_i is a cause of X_j ($X_i \rightarrow X_j$), iff there exist values of X_i and X_j such that $p(x_j|\text{do}\{x_i\}) \neq p(x_j)$.

The chain	The fork	The collider
$X_1 \rightarrow X_2 \rightarrow X_3$	$X_1 \leftarrow X_2 \rightarrow X_3$	$X_1 \rightarrow X_2 \leftarrow X_3$
$X_1 \not\perp\!\!\!\perp X_3$	$X_1 \not\perp\!\!\!\perp X_3$	$X_1 \perp\!\!\!\perp X_3$
$X_1 \perp\!\!\!\perp X_3 \mid X_2$	$X_1 \perp\!\!\!\perp X_3 \mid X_2$	$X_1 \not\perp\!\!\!\perp X_3 \mid X_2$

- Causal Markov condition: Independence relations implied by a directed acyclic graph (DAG) are encoded in every $p(\mathbf{X})$ generated by this DAG.
- Faithfulness: $p(\mathbf{X})$ contains no additional independence relations beyond those implied by its generating DAG.

Causal Terminology Revisited

Causal Terminology Revisited

Hippocampal activity in this study was **correlated** *with amygdala activity, supporting the view that the amygdala* **enhances** *explicit memory by* **modulating** *activity in the hippocampus.*

(Anonymous author, *Trends in Cognitive Sciences*, 2001)

Causal Terminology Revisited

Hippocampal activity in this study was **correlated** *with amygdala activity, supporting the view that the amygdala* **enhances** *explicit memory by* **modulating** *activity in the hippocampus.*

(Anonymous author, *Trends in Cognitive Sciences*, 2001)

$$\text{HC} \not\perp\!\!\!\perp \text{AM} \Rightarrow \text{AM} \rightarrow \text{HC} \rightarrow \text{EM}$$

Causal Terminology Revisited

Hippocampal activity in this study was **correlated** *with amygdala activity, supporting the view that the amygdala* **enhances** *explicit memory by* **modulating** *activity in the hippocampus.*

(Anonymous author, *Trends in Cognitive Sciences*, 2001)

$$\text{HC} \not\perp\!\!\!\perp \text{AM} \Rightarrow \text{AM} \rightarrow \text{HC} \rightarrow \text{EM}$$

We tested [...] whether pre-stimulus alpha oscillations **measured** *with electroencephalography (EEG)* **influence** *the encoding of items into working memory.*

(Anonymous authors, *Journal of Neuroscience*, 2014)

Causal Terminology Revisited

Hippocampal activity in this study was **correlated** *with amygdala activity, supporting the view that the amygdala* **enhances** *explicit memory by* **modulating** *activity in the hippocampus.*

(Anonymous author, *Trends in Cognitive Sciences*, 2001)

$$\mathrm{HC} \not\perp\!\!\!\perp \mathrm{AM} \Rightarrow \mathrm{AM} \rightarrow \mathrm{HC} \rightarrow \mathrm{EM}$$

We tested [...] whether pre-stimulus alpha oscillations **measured** *with electroencephalography (EEG)* **influence** *the encoding of items into working memory.*

(Anonymous authors, *Journal of Neuroscience*, 2014)

$$\alpha \not\perp\!\!\!\perp \mathrm{WM} \Rightarrow \alpha \rightarrow \mathrm{WM}$$

Causal Inference in Encoding/Decoding Models

Causal Inference in Encoding/Decoding Models

Encoding models: $p(\mathbf{X}|C)$

Causal Inference in Encoding/Decoding Models

Encoding models: $p(\mathbf{X}|C)$

- Does a brain-state feature X_i change across experimental conditions?

Causal Inference in Encoding/Decoding Models

Encoding models: $p(\mathbf{X}|C)$

- Does a brain-state feature X_i change across experimental conditions?

$$H0: \; C \perp\!\!\!\perp X_i$$

Causal Inference in Encoding/Decoding Models

Encoding models: $p(\mathbf{X}|C)$

- Does a brain-state feature X_i change across experimental conditions?
- Reject $H0: \; C \perp\!\!\!\perp X_i \Rightarrow X_i$ is a relevant feature

Causal Inference in Encoding/Decoding Models

Encoding models: $p(\mathbf{X}|C)$

- Does a brain-state feature X_i change across experimental conditions?
- Reject $H0: \; C \perp\!\!\!\perp X_i \Rightarrow X_i$ is a relevant feature

Decoding models: $p(C|\mathbf{X})$

Causal Inference in Encoding/Decoding Models

Encoding models: $p(\mathbf{X}|C)$

- Does a brain-state feature X_i change across experimental conditions?
- Reject $H0: \; C \perp\!\!\!\perp X_i \Rightarrow X_i$ is a relevant feature

Decoding models: $p(C|\mathbf{X})$

- Does a brain-state feature help in decoding the experimental condition?

Causal Inference in Encoding/Decoding Models

Encoding models: $p(\mathbf{X}|C)$

- Does a brain-state feature X_i change across experimental conditions?
- Reject $H0: \; C \perp\!\!\!\perp X_i \Rightarrow X_i$ is a relevant feature

Decoding models: $p(C|\mathbf{X})$

- Does a brain-state feature help in decoding the experimental condition?

$$H0: \; C \perp\!\!\!\perp X_i|\mathbf{X}\backslash X_i$$

Causal Inference in Encoding/Decoding Models

Encoding models: $p(\mathbf{X}|C)$

- Does a brain-state feature X_i change across experimental conditions?
- Reject $H0: \; C \perp\!\!\!\perp X_i \Rightarrow X_i$ is a relevant feature

Decoding models: $p(C|\mathbf{X})$

- Does a brain-state feature help in decoding the experimental condition?
- Reject $H0: \; C \perp\!\!\!\perp X_i|\mathbf{X}\backslash X_i \Rightarrow X_i$ is a relevant feature

Causal Inference in Encoding/Decoding Models

Encoding models: $p(\mathbf{X}|C)$

- Does a brain-state feature X_i change across experimental conditions?
- Reject $H0: \; C \perp\!\!\!\perp X_i \Rightarrow X_i$ is a relevant feature

Decoding models: $p(C|\mathbf{X})$

- Does a brain-state feature help in decoding the experimental condition?
- Reject $H0: \; C \perp\!\!\!\perp X_i|\mathbf{X}\backslash X_i \Rightarrow X_i$ is a relevant feature

$$\text{Feature relevance} \Rightarrow C \not\perp\!\!\!\perp X_i \; / \; C \not\perp\!\!\!\perp X_i|\mathbf{X}\backslash X_i \Rightarrow \text{causal structure}$$

Outline

1. Causal Interpretation Rules for Individual Encoding/Decoding Models
2. Causal Interpretation Rules for Joint Encoding/Decoding Models

Outline

1. Causal Interpretation Rules for Individual Encoding/Decoding Models
2. Causal Interpretation Rules for Joint Encoding/Decoding Models

Causal Interpretation Rules for Individual Models

Experimental setting		Feature X_i relevant? Encoding	Feature X_i relevant? Decoding	Causal interpretation
Task		$\times$		
		$\surd$		
			$\times$	
			$\surd$	
Behaviour		$\times$		
		$\surd$		
			$\times$	
			$\surd$	

Causal Interpretation Rules for Individual Models

Experimental setting		Feature X_i relevant? Encoding	Feature X_i relevant? Decoding	Causal interpretation
Task		×		$T \perp\!\!\!\perp X_i$
Task		√		
Task			×	
Task			√	
Behaviour		×		
Behaviour		√		
Behaviour			×	
Behaviour			√	

Causal Interpretation Rules for Individual Models

Experimental setting		Feature X_i relevant? Encoding	Feature X_i relevant? Decoding	Causal interpretation
Task		×		$T \perp\!\!\!\perp X_i \Rightarrow X_i$ is no effect of T
Task		√		
Task			×	
Task			√	
Behaviour		×		
Behaviour		√		
Behaviour			×	
Behaviour			√	

Causal Interpretation Rules for Individual Models

Experimental setting		Feature X_i relevant? Encoding	Feature X_i relevant? Decoding	Causal interpretation
Task		$\times$		$T \perp\!\!\!\perp X_i \Rightarrow X_i$ is no effect of T
Task		$\surd$		
Task			$\times$	$T \perp\!\!\!\perp X_i \vert \mathbf{X} \backslash X_i$
Task			$\surd$	
Behaviour		$\times$		
Behaviour		$\surd$		
Behaviour			$\times$	
Behaviour			$\surd$	

Causal Interpretation Rules for Individual Models

Experimental setting		Feature X_i relevant? Encoding	Feature X_i relevant? Decoding	Causal interpretation
Task		×		$T \perp\!\!\!\perp X_i \Rightarrow X_i$ is no effect of T
Task		√		
Task			×	$T \perp\!\!\!\perp X_i \vert \mathbf{X} \backslash X_i$
Task			√	
Behaviour		×		
Behaviour		√		
Behaviour			×	
Behaviour			√	

T X_i vs. $T \rightarrow X_j \rightarrow X_i$

Causal Interpretation Rules for Individual Models

Experimental setting		Feature X_i relevant? Encoding	Feature X_i relevant? Decoding	Causal interpretation
Task		×		$T \perp\!\!\!\perp X_i \Rightarrow X_i$ is no effect of T
Task		√		
Task			×	$T \perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i \Rightarrow$ inconclusive
Task			√	
Behaviour		×		
Behaviour		√		
Behaviour			×	
Behaviour			√	

T X_i vs. $T \rightarrow X_j \rightarrow X_i$

Causal Interpretation Rules for Individual Models

Experimental setting		Feature X_i relevant? Encoding	Feature X_i relevant? Decoding	Causal interpretation
Task		×		$T \perp\!\!\!\perp X_i \Rightarrow X_i$ is no effect of T
Task		√		$T \not\perp\!\!\!\perp X_i$
Task			×	$T \perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i \Rightarrow$ inconclusive
Task			√	
Behaviour		×		
Behaviour		√		
Behaviour			×	
Behaviour			√	

Causal Interpretation Rules for Individual Models

Experimental setting		Feature X_i relevant? Encoding	Feature X_i relevant? Decoding	Causal interpretation
Task		×		$T \perp\!\!\!\perp X_i \Rightarrow X_i$ is no effect of T
Task		√		$T \not\perp\!\!\!\perp X_i \Rightarrow X_i$ is an effect of T
Task			×	$T \perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i \Rightarrow$ inconclusive
Task			√	
Behaviour		×		
Behaviour		√		
Behaviour			×	
Behaviour			√	

Causal Interpretation Rules for Individual Models

Experimental setting		Feature X_i relevant? Encoding	Feature X_i relevant? Decoding	Causal interpretation
Task		×		$T \perp\!\!\!\perp X_i \Rightarrow X_i$ is no effect of T
Task		√		$T \not\perp\!\!\!\perp X_i \Rightarrow X_i$ is an effect of T
Task			×	$T \perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i \Rightarrow$ inconclusive
Task			√	$T \not\perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i$
Behaviour		×		
Behaviour		√		
Behaviour			×	
Behaviour			√	

Causal Interpretation Rules for Individual Models

Experimental setting		Feature X_i relevant? Encoding	Feature X_i relevant? Decoding	Causal interpretation
Task		×		$T \perp\!\!\!\perp X_i \Rightarrow X_i$ is no effect of T
Task		√		$T \not\perp\!\!\!\perp X_i \Rightarrow X_i$ is an effect of T
Task			×	$T \perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i \Rightarrow$ inconclusive
Task			√	$T \not\perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i$
Behaviour		×		
Behaviour		√		
Behaviour			×	
Behaviour			√	

$$T \rightarrow X_i \text{ or } T \rightarrow X_j \leftarrow X_i$$

Causal Interpretation Rules for Individual Models

Experimental setting		Feature X_i relevant? Encoding	Feature X_i relevant? Decoding	Causal interpretation
Task		$\times$		$T \perp\!\!\!\perp X_i \Rightarrow X_i$ is no effect of T
Task		$\surd$		$T \not\perp\!\!\!\perp X_i \Rightarrow X_i$ is an effect of T
Task			$\times$	$T \perp\!\!\!\perp X_i \vert \mathbf{X} \backslash X_i \Rightarrow$ inconclusive
Task			$\surd$	$T \not\perp\!\!\!\perp X_i \vert \mathbf{X} \backslash X_i \Rightarrow$ inconclusive
Behaviour		$\times$		
Behaviour		$\surd$		
Behaviour			$\times$	
Behaviour			$\surd$	

$$T \rightarrow X_i \text{ or } T \rightarrow X_j \leftarrow X_i$$

Causal Interpretation Rules for Individual Models

Experimental setting		Feature X_i relevant? Encoding	Feature X_i relevant? Decoding	Causal interpretation
Task		$\times$		$T \perp\!\!\!\perp X_i \Rightarrow X_i$ is no effect of T
Task		$\surd$		$T \not\perp\!\!\!\perp X_i \Rightarrow X_i$ is an effect of T
Task			$\times$	$T \perp\!\!\!\perp X_i \vert \mathbf{X} \backslash X_i \Rightarrow$ inconclusive
Task			$\surd$	$T \not\perp\!\!\!\perp X_i \vert \mathbf{X} \backslash X_i \Rightarrow$ inconclusive
Behaviour		$\times$		$B \perp\!\!\!\perp X_i$
Behaviour		$\surd$		
Behaviour			$\times$	
Behaviour			$\surd$	

Causal Interpretation Rules for Individual Models

Experimental setting		Feature X_i relevant? Encoding	Feature X_i relevant? Decoding	Causal interpretation
Task		×		$T \perp\!\!\!\perp X_i \Rightarrow X_i$ is no effect of T
Task		√		$T \not\perp\!\!\!\perp X_i \Rightarrow X_i$ is an effect of T
Task			×	$T \perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i \Rightarrow$ inconclusive
Task			√	$T \not\perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i \Rightarrow$ inconclusive
Behaviour		×		$B \perp\!\!\!\perp X_i \Rightarrow X_i$ is no cause of B
Behaviour		√		
Behaviour			×	
Behaviour			√	

Causal Interpretation Rules for Individual Models

Experimental setting		Feature X_i relevant? Encoding	Feature X_i relevant? Decoding	Causal interpretation
Task		×		$T \perp\!\!\!\perp X_i \Rightarrow X_i$ is no effect of T
Task		√		$T \not\perp\!\!\!\perp X_i \Rightarrow X_i$ is an effect of T
Task			×	$T \perp\!\!\!\perp X_i \vert \mathbf{X} \backslash X_i \Rightarrow$ inconclusive
Task			√	$T \not\perp\!\!\!\perp X_i \vert \mathbf{X} \backslash X_i \Rightarrow$ inconclusive
Behaviour		×		$B \perp\!\!\!\perp X_i \Rightarrow X_i$ is no cause of B
Behaviour		√		
Behaviour			×	$B \perp\!\!\!\perp X_i \vert \mathbf{X} \backslash X_i$
Behaviour			√	

Causal Interpretation Rules for Individual Models

Experimental setting		Feature X_i relevant? Encoding	Feature X_i relevant? Decoding	Causal interpretation
Task		×		$T \perp\!\!\!\perp X_i \Rightarrow X_i$ is no effect of T
Task		√		$T \not\perp\!\!\!\perp X_i \Rightarrow X_i$ is an effect of T
Task			×	$T \perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i \Rightarrow$ inconclusive
Task			√	$T \not\perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i \Rightarrow$ inconclusive
Behaviour		×		$B \perp\!\!\!\perp X_i \Rightarrow X_i$ is no cause of B
Behaviour		√		
Behaviour			×	$B \perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i$
Behaviour			√	

X_i $\quad$ B vs. $X_i \rightarrow X_j \rightarrow B$

Causal Interpretation Rules for Individual Models

Experimental setting		Feature X_i relevant? Encoding	Feature X_i relevant? Decoding	Causal interpretation
Task		×		$T \perp\!\!\!\perp X_i \Rightarrow X_i$ is no effect of T
Task		√		$T \not\perp\!\!\!\perp X_i \Rightarrow X_i$ is an effect of T
Task			×	$T \perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i \Rightarrow$ inconclusive
Task			√	$T \not\perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i \Rightarrow$ inconclusive
Behaviour		×		$B \perp\!\!\!\perp X_i \Rightarrow X_i$ is no cause of B
Behaviour		√		
Behaviour			×	$B \perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i \Rightarrow$ inconclusive
Behaviour			√	

X_i B vs. $X_i \rightarrow X_j \rightarrow B$

Causal Interpretation Rules for Individual Models

Experimental setting		Feature X_i relevant? Encoding	Feature X_i relevant? Decoding	Causal interpretation
Task		×		$T \perp\!\!\!\perp X_i \Rightarrow X_i$ is no effect of T
Task		√		$T \not\perp\!\!\!\perp X_i \Rightarrow X_i$ is an effect of T
Task			×	$T \perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i \Rightarrow$ inconclusive
Task			√	$T \not\perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i \Rightarrow$ inconclusive
Behaviour		×		$B \perp\!\!\!\perp X_i \Rightarrow X_i$ is no cause of B
Behaviour		√		$B \not\perp\!\!\!\perp X_i$
Behaviour			×	$B \perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i \Rightarrow$ inconclusive
Behaviour			√	

Causal Interpretation Rules for Individual Models

Experimental setting		Feature X_i relevant? Encoding	Feature X_i relevant? Decoding	Causal interpretation
Task		×		$T \perp\!\!\!\perp X_i \Rightarrow X_i$ is no effect of T
Task		√		$T \not\perp\!\!\!\perp X_i \Rightarrow X_i$ is an effect of T
Task			×	$T \perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i \Rightarrow$ inconclusive
Task			√	$T \not\perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i \Rightarrow$ inconclusive
Behaviour		×		$B \perp\!\!\!\perp X_i \Rightarrow X_i$ is no cause of B
Behaviour		√		$B \not\perp\!\!\!\perp X_i$
Behaviour			×	$B \perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i \Rightarrow$ inconclusive
Behaviour			√	

$$X_i \rightarrow B \text{ or } X_i \leftarrow X_j \rightarrow B$$

Causal Interpretation Rules for Individual Models

Experimental setting		Feature X_i relevant? Encoding	Feature X_i relevant? Decoding	Causal interpretation
	Task	×		$T \perp\!\!\!\perp X_i \Rightarrow X_i$ is no effect of T
		√		$T \not\perp\!\!\!\perp X_i \Rightarrow X_i$ is an effect of T
			×	$T \perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i \Rightarrow$ inconclusive
			√	$T \not\perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i \Rightarrow$ inconclusive
	Behaviour	×		$B \perp\!\!\!\perp X_i \Rightarrow X_i$ is no cause of B
		√		$B \not\perp\!\!\!\perp X_i \Rightarrow$ inconclusive
			×	$B \perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i \Rightarrow$ inconclusive
			√	

$$X_i \rightarrow B \text{ or } X_i \leftarrow X_j \rightarrow B$$

Causal Interpretation Rules for Individual Models

Experimental setting		Feature X_i relevant? Encoding	Feature X_i relevant? Decoding	Causal interpretation
Task		$\times$		$T \perp\!\!\!\perp X_i \Rightarrow X_i$ is no effect of T
		$\surd$		$T \not\perp\!\!\!\perp X_i \Rightarrow X_i$ is an effect of T
			$\times$	$T \perp\!\!\!\perp X_i \vert \mathbf{X} \backslash X_i \Rightarrow$ inconclusive
			$\surd$	$T \not\perp\!\!\!\perp X_i \vert \mathbf{X} \backslash X_i \Rightarrow$ inconclusive
Behaviour		$\times$		$B \perp\!\!\!\perp X_i \Rightarrow X_i$ is no cause of B
		$\surd$		$B \not\perp\!\!\!\perp X_i \Rightarrow$ inconclusive
			$\times$	$B \perp\!\!\!\perp X_i \vert \mathbf{X} \backslash X_i \Rightarrow$ inconclusive
			$\surd$	$B \not\perp\!\!\!\perp X_i \vert \mathbf{X} \backslash X_i$

Causal Interpretation Rules for Individual Models

Experimental setting		Feature X_i relevant? Encoding	Feature X_i relevant? Decoding	Causal interpretation
Task		$\times$		$T \perp\!\!\!\perp X_i \Rightarrow X_i$ is no effect of T
Task		$\surd$		$T \not\perp\!\!\!\perp X_i \Rightarrow X_i$ is an effect of T
Task			$\times$	$T \perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i \Rightarrow$ inconclusive
Task			$\surd$	$T \not\perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i \Rightarrow$ inconclusive
Behaviour		$\times$		$B \perp\!\!\!\perp X_i \Rightarrow X_i$ is no cause of B
Behaviour		$\surd$		$B \not\perp\!\!\!\perp X_i \Rightarrow$ inconclusive
Behaviour			$\times$	$B \perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i \Rightarrow$ inconclusive
Behaviour			$\surd$	$B \not\perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i$

$$X_i \rightarrow B \text{ or } X_i \leftarrow H \rightarrow X_j$$
$$\downarrow$$
$$B$$

Causal Interpretation Rules for Individual Models

Experimental setting		Feature X_i relevant? Encoding	Feature X_i relevant? Decoding	Causal interpretation
Task		×		$T \perp\!\!\!\perp X_i \Rightarrow X_i$ is no effect of T
Task		√		$T \not\perp\!\!\!\perp X_i \Rightarrow X_i$ is an effect of T
Task			×	$T \perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i \Rightarrow$ inconclusive
Task			√	$T \not\perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i \Rightarrow$ inconclusive
Behaviour		×		$B \perp\!\!\!\perp X_i \Rightarrow X_i$ is no cause of B
Behaviour		√		$B \not\perp\!\!\!\perp X_i \Rightarrow$ inconclusive
Behaviour			×	$B \perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i \Rightarrow$ inconclusive
Behaviour			√	$B \not\perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i \Rightarrow$ inconclusive

$$X_i \rightarrow B \text{ or } X_i \leftarrow H \rightarrow X_j$$
$$\downarrow$$
$$B$$

Causal Interpretation Rules for Individual Models

Experimental setting		Feature X_i relevant? Encoding	Feature X_i relevant? Decoding	Causal interpretation
Task		×		$T \perp\!\!\!\perp X_i \Rightarrow X_i$ is no effect of T
		√		$T \not\perp\!\!\!\perp X_i \Rightarrow X_i$ is an effect of T
			×	$T \perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i \Rightarrow$ inconclusive
			√	$T \not\perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i \Rightarrow$ inconclusive
Behaviour		×		$B \perp\!\!\!\perp X_i \Rightarrow X_i$ is no cause of B
		√		$B \not\perp\!\!\!\perp X_i \Rightarrow$ inconclusive
			×	$B \perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i \Rightarrow$ inconclusive
			√	$B \not\perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i \Rightarrow$ inconclusive

We tested [...] whether pre-stimulus alpha oscillations **measured** *with electroencephalography (EEG)* **influence** *the encoding of items into working memory.* (Anonymous authors, *Journal of Neuroscience*, 2014)

Causal Interpretation Rules for Individual Models

Experimental setting		Feature X_i relevant? Encoding	Feature X_i relevant? Decoding	Causal interpretation
Task		×		$T \perp\!\!\!\perp X_i \Rightarrow X_i$ is no effect of T
		√		$T \not\perp\!\!\!\perp X_i \Rightarrow X_i$ is an effect of T
			×	$T \perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i \Rightarrow$ inconclusive
			√	$T \not\perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i \Rightarrow$ inconclusive
Behaviour		×		$B \perp\!\!\!\perp X_i \Rightarrow X_i$ is no cause of B
		√		$B \not\perp\!\!\!\perp X_i \Rightarrow$ inconclusive
			×	$B \perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i \Rightarrow$ inconclusive
			√	$B \not\perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i \Rightarrow$ inconclusive

We tested [...] whether pre-stimulus alpha oscillations **measured** *with electroencephalography (EEG)* **influence** *the encoding of items into working memory.* (Anonymous authors, *Journal of Neuroscience*, 2014)

$$\alpha \not\perp\!\!\!\perp \text{WM} \Rightarrow \alpha \rightarrow \text{WM}$$

Causal Interpretation Rules for Individual Models

Experimental setting		Feature X_i relevant? Encoding	Feature X_i relevant? Decoding	Causal interpretation
Task		×		$T \perp\!\!\!\perp X_i \Rightarrow X_i$ is no effect of T
Task		√		$T \not\perp\!\!\!\perp X_i \Rightarrow X_i$ is an effect of T
Task			×	$T \perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i \Rightarrow$ inconclusive
Task			√	$T \not\perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i \Rightarrow$ inconclusive
Behaviour		×		$B \perp\!\!\!\perp X_i \Rightarrow X_i$ is no cause of B
Behaviour		√		$B \not\perp\!\!\!\perp X_i \Rightarrow$ inconclusive
Behaviour			×	$B \perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i \Rightarrow$ inconclusive
Behaviour			√	$B \not\perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i \Rightarrow$ inconclusive

Outline

1 Causal Interpretation Rules for Individual Encoding/Decoding Models

2 Causal Interpretation Rules for Joint Encoding/Decoding Models

Causal Interpretation Rules for Joint Models

Experimental setting		Feature X_i relevant? Encoding	Feature X_i relevant? Decoding	Causal interpretation
Task		$\times$	$\times$	
Task		$\surd$	$\times$	
Task		$\times$	$\surd$	
Task		$\surd$	$\surd$	
Behaviour		$\times$	$\times$	
Behaviour		$\surd$	$\times$	
Behaviour		$\times$	$\surd$	
Behaviour		$\surd$	$\surd$	

Causal Interpretation Rules for Joint Models

Experimental setting		Feature X_i relevant? Encoding	Decoding	Causal interpretation
Task		$\times$	$\times$	
		$\surd$	$\times$	$T \not\perp\!\!\!\perp X_i \;\&\; T \perp\!\!\!\perp X_i \vert \mathbf{X} \backslash X_i$
		$\times$	$\surd$	
		$\surd$	$\surd$	
Behaviour		$\times$	$\times$	
		$\surd$	$\times$	
		$\times$	$\surd$	
		$\surd$	$\surd$	

Causal Interpretation Rules for Joint Models

Experimental setting		Feature X_i relevant? Encoding	Decoding	Causal interpretation
Task		$\times$	$\times$	
		$\surd$	$\times$	$T \not\perp\!\!\!\perp X_i \;\&\; T \perp\!\!\!\perp X_i \vert \mathbf{X} \backslash X_i$
		$\times$	$\surd$	
		$\surd$	$\surd$	
Behaviour		$\times$	$\times$	
		$\surd$	$\times$	
		$\times$	$\surd$	
		$\surd$	$\surd$	

$$T \rightarrow X_j \rightarrow X_i$$

Causal Interpretation Rules for Joint Models

Experimental setting		Feature X_i relevant? Encoding	Decoding	Causal interpretation
Task		$\times$	$\times$	
		$\surd$	$\times$	X_i is an indirect effect of T
		$\times$	$\surd$	
		$\surd$	$\surd$	
Behaviour		$\times$	$\times$	
		$\surd$	$\times$	
		$\times$	$\surd$	
		$\surd$	$\surd$	

$$T \rightarrow X_j \rightarrow X_i$$

Causal Interpretation Rules for Joint Models

Experimental setting		Feature X_i relevant? Encoding	Feature X_i relevant? Decoding	Causal interpretation
Task		$\times$	$\times$	
		$\surd$	$\times$	X_i is an indirect effect of T
		$\times$	$\surd$	$T \perp\!\!\!\perp X_i$ & $T \not\perp\!\!\!\perp X_i \vert \mathbf{X} \backslash X_i$
		$\surd$	$\surd$	
Behaviour		$\times$	$\times$	
		$\surd$	$\times$	
		$\times$	$\surd$	
		$\surd$	$\surd$	

Causal Interpretation Rules for Joint Models

		Feature X_i relevant? Encoding	Decoding	Causal interpretation
Experimental setting	Task	×	×	
		√	×	X_i is an indirect effect of T
		×	√	$T \perp\!\!\!\perp X_i$ & $T \not\perp\!\!\!\perp X_i \vert \mathbf{X} \backslash X_i$
		√	√	
	Behaviour	×	×	
		√	×	
		×	√	
		√	√	

$$T \rightarrow X_j \leftarrow X_i \text{ or } T \rightarrow X_j \leftarrow H \rightarrow X_i$$

Causal Interpretation Rules for Joint Models

Experimental setting		Feature X_i relevant? Encoding	Feature X_i relevant? Decoding	Causal interpretation
Task		$\times$	$\times$	
		$\surd$	$\times$	X_i is an indirect effect of T
		$\times$	$\surd$	X_i provides context
		$\surd$	$\surd$	
Behaviour		$\times$	$\times$	
		$\surd$	$\times$	
		$\times$	$\surd$	
		$\surd$	$\surd$	

$$T \rightarrow X_j \leftarrow X_i \text{ or } T \rightarrow X_j \leftarrow H \rightarrow X_i$$

Causal Interpretation Rules for Joint Models

Experimental setting		Feature X_i relevant? Encoding	Feature X_i relevant? Decoding	Causal interpretation
Task		$\times$	$\times$	X_i is no effect of T
		$\surd$	$\times$	X_i is an indirect effect of T
		$\times$	$\surd$	X_i provides context
		$\surd$	$\surd$	X_i is an effect of T
Behaviour		$\times$	$\times$	X_i is no cause of B
		$\surd$	$\times$	X_i is no direct cause of B
		$\times$	$\surd$	X_i provides context
		$\surd$	$\surd$	inconclusive

Causal Interpretation Rules for Joint Models

Experimental setting		Feature X_i relevant? Encoding	Feature X_i relevant? Decoding	Causal interpretation
Task		×	×	X_i is no effect of T
		√	×	X_i is an indirect effect of T
		×	√	X_i provides context
		√	√	X_i is an effect of T
Behaviour		×	×	X_i is no cause of B
		√	×	X_i is no direct cause of B
		×	√	X_i provides context
		√	√	inconclusive

Causal Interpretation Rules for Joint Models

		Feature X_i relevant? Encoding	Decoding	Causal interpretation
Experimental setting	Task	×	×	X_i is no effect of T
		√	×	X_i is an indirect effect of T
		×	√	X_i provides context
		√	√	X_i is an effect of T
	Behaviour	×	×	X_i is no cause of B
		√	×	X_i is no direct cause of B
		×	√	X_i provides context
		√	√	inconclusive

Hippocampal activity in this study was **correlated** *with amygdala activity, supporting the view that the amygdala* **enhances** *explicit memory by* **modulating** *activity in the hippocampus. (Anonymous author, Trends in Cognitive Sciences, 2001)*

Causal Interpretation Rules for Joint Models

		Feature X_i relevant? Encoding	Decoding	Causal interpretation
Experimental setting	Task	×	×	X_i is no effect of T
		√	×	X_i is an indirect effect of T
		×	√	X_i provides context
		√	√	X_i is an effect of T
	Behaviour	×	×	X_i is no cause of B
		√	×	X_i is no direct cause of B
		×	√	X_i provides context
		√	√	inconclusive

Hippocampal activity in this study was **correlated** *with amygdala activity, supporting the view that the amygdala* **enhances** *explicit memory by* **modulating** *activity in the hippocampus. (Anonymous author, Trends in Cognitive Sciences, 2001)*

$$\text{HC} \not\perp\!\!\!\perp \text{AM} \Rightarrow \text{AM} \rightarrow \text{HC} \rightarrow \text{EM}$$

Causal Interpretation Rules for Joint Models

		Feature X_i relevant? Encoding	Decoding	Causal interpretation
Experimental setting	Task	×	×	X_i is no effect of T
		√	×	X_i is an indirect effect of T
		×	√	X_i provides context
		√	√	X_i is an effect of T
	Behaviour	×	×	X_i is no cause of B
		√	×	X_i is no direct cause of B
		×	√	X_i provides context
		√	√	inconclusive

Hippocampal activity in this study was **correlated** *with amygdala activity, supporting the view that the amygdala* **enhances** *explicit memory by* **modulating** *activity in the hippocampus. (Anonymous author, Trends in Cognitive Sciences, 2001)*

$$\mathrm{AM} \not\perp\!\!\!\perp \mathrm{EM} \;\&\; \mathrm{AM} \perp\!\!\!\perp \mathrm{EM} | \mathrm{HC} \Leftarrow \mathrm{AM} \rightarrow \mathrm{HC} \rightarrow \mathrm{EM}$$

Wrapping Up

Wrapping Up

Caveats:

Wrapping Up

Caveats:

- The probability of type II errors can not be quantified in the same manner as the probability of type I errors.

Wrapping Up

Caveats:

- The probability of type II errors can not be quantified in the same manner as the probability of type I errors.
- The assumption of faithfulness is presently untestable.

Wrapping Up

Caveats:

- The probability of type II errors can not be quantified in the same manner as the probability of type I errors.
- The assumption of faithfulness is presently untestable.
- Permutation-based relevance tests in decoding models are biased towards dependence (Strobl et al., *BMC Bioinformatics*, 2008)

Wrapping Up

Caveats:

- The probability of type II errors can not be quantified in the same manner as the probability of type I errors.
- The assumption of faithfulness is presently untestable.
- Permutation-based relevance tests in decoding models are biased towards dependence (Strobl et al., *BMC Bioinformatics*, 2008)
- Unbiased conditional independence tests are hard (Zhang et al., *UAI*, 2011)

Wrapping Up

Caveats:

- The probability of type II errors can not be quantified in the same manner as the probability of type I errors.
- The assumption of faithfulness is presently untestable.
- Permutation-based relevance tests in decoding models are biased towards dependence (Strobl et al., *BMC Bioinformatics*, 2008)
- Unbiased conditional independence tests are hard (Zhang et al., *UAI*, 2011)

Take home message:

Wrapping Up

Caveats:

- The probability of type II errors can not be quantified in the same manner as the probability of type I errors.
- The assumption of faithfulness is presently untestable.
- Permutation-based relevance tests in decoding models are biased towards dependence (Strobl et al., *BMC Bioinformatics*, 2008)
- Unbiased conditional independence tests are hard (Zhang et al., *UAI*, 2011)

Take home message:

- If you don't like causal inference, don't use causal terminology.

Wrapping Up

Caveats:

- The probability of type II errors can not be quantified in the same manner as the probability of type I errors.
- The assumption of faithfulness is presently untestable.
- Permutation-based relevance tests in decoding models are biased towards dependence (Strobl et al., *BMC Bioinformatics*, 2008)
- Unbiased conditional independence tests are hard (Zhang et al., *UAI*, 2011)

Take home message:

- If you don't like causal inference, don't use causal terminology.
- If you use causal terminology, make sure that

Wrapping Up

Caveats:

- The probability of type II errors can not be quantified in the same manner as the probability of type I errors.
- The assumption of faithfulness is presently untestable.
- Permutation-based relevance tests in decoding models are biased towards dependence (Strobl et al., *BMC Bioinformatics*, 2008)
- Unbiased conditional independence tests are hard (Zhang et al., *UAI*, 2011)

Take home message:

- If you don't like causal inference, don't use causal terminology.
- If you use causal terminology, make sure that
  - your conclusions are supported by empirical data

Wrapping Up

Caveats:

- The probability of type II errors can not be quantified in the same manner as the probability of type I errors.
- The assumption of faithfulness is presently untestable.
- Permutation-based relevance tests in decoding models are biased towards dependence (Strobl et al., *BMC Bioinformatics*, 2008)
- Unbiased conditional independence tests are hard (Zhang et al., *UAI*, 2011)

Take home message:

- If you don't like causal inference, don't use causal terminology.
- If you use causal terminology, make sure that
  - your conclusions are supported by empirical data
  - you are explicit about inherent assumptions

Acknowledgements

People:

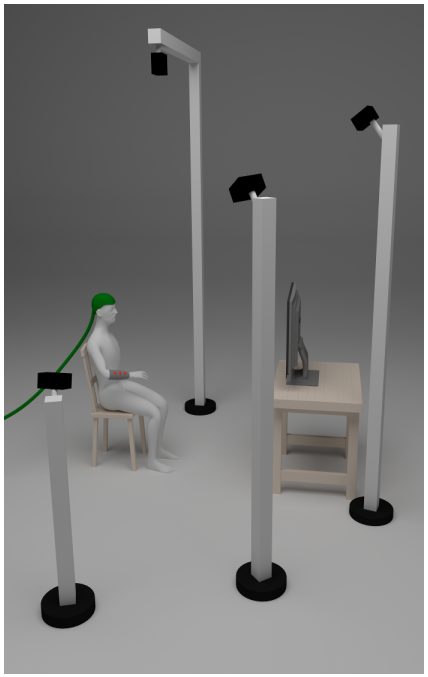
- Sebastian Weichwald (UCL)
- Timm Meyer (MPI-IS)
- Ozan Özdenizci (Sabanci University)
- Bernhard Schölkopf (MPI-IS)
- Tonio Ball (University of Freiburg)

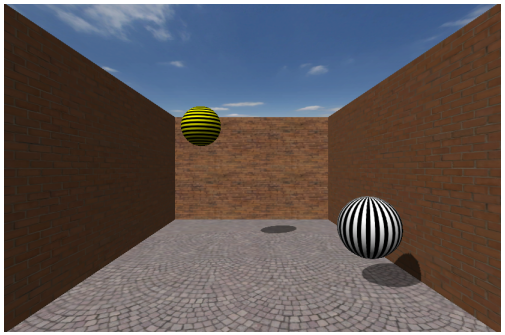
Publications:

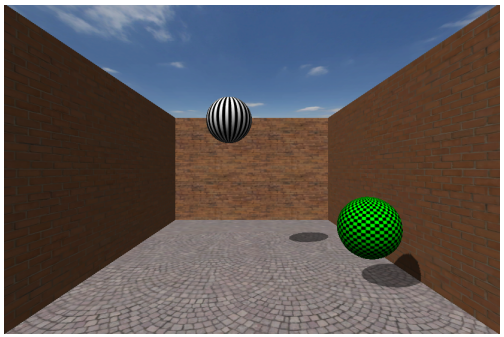
- Weichwald et al., Causal interpretation rules for encoding and decoding models in neuroimaging. *NeuroImage*, 2015.
- Weichwald et al., Causal and anti-causal learning in pattern recognition for neuroimaging. *PRNI*, 2014.

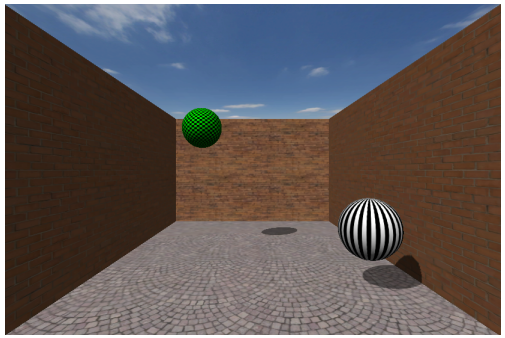
`http://brain-computer-interfaces.net`

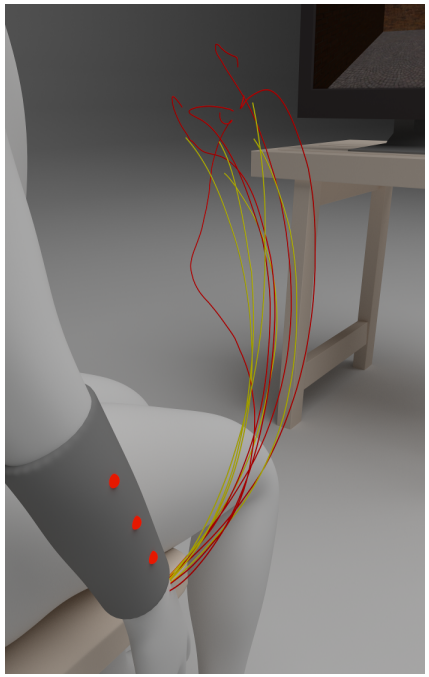
Experimental Setup



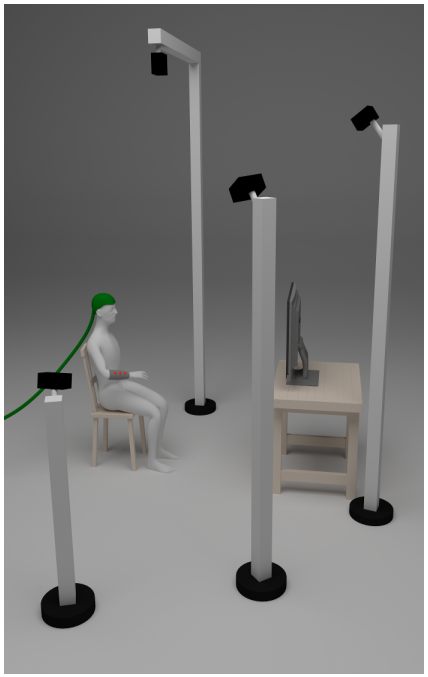


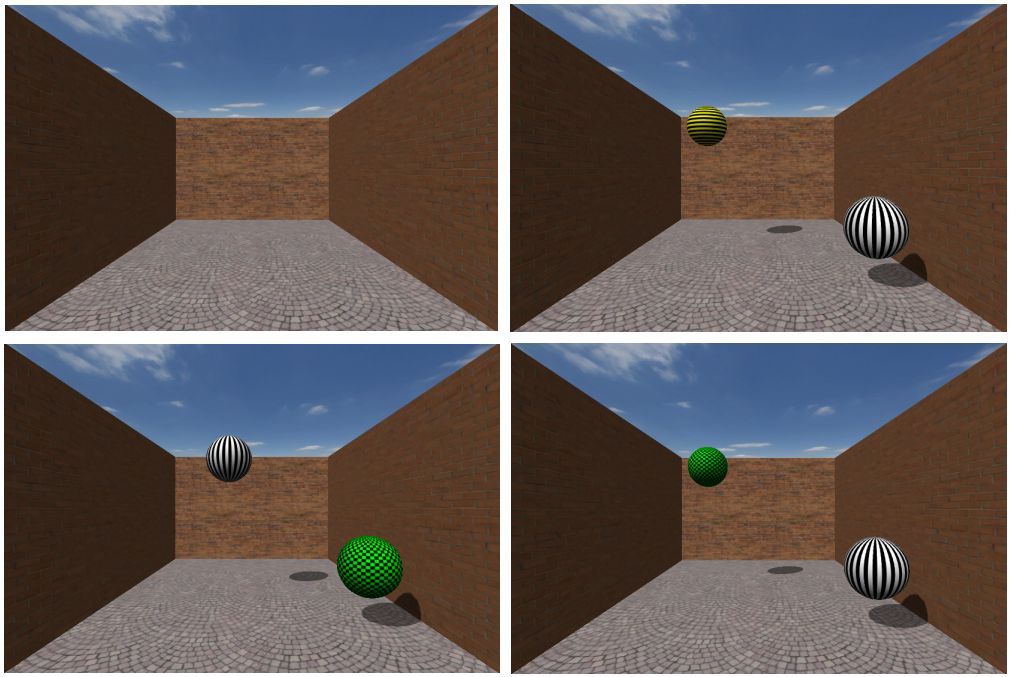


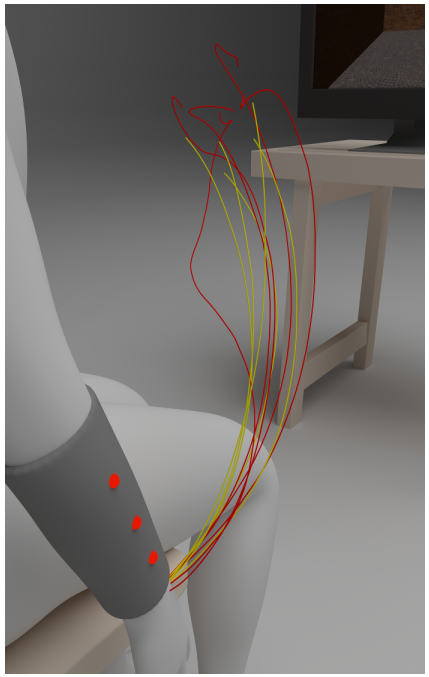


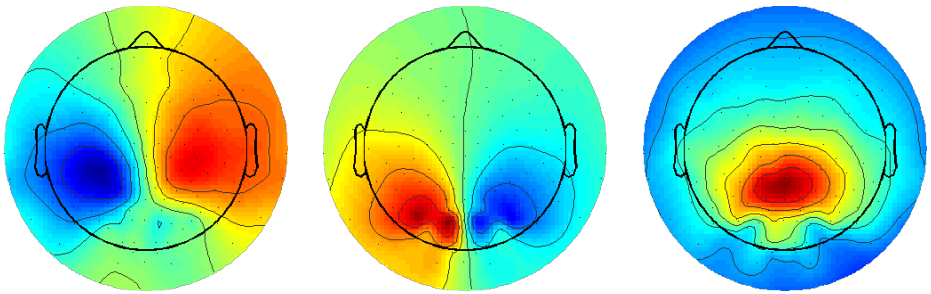


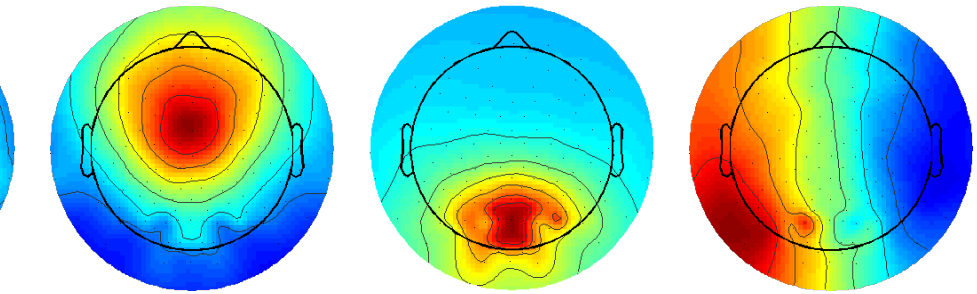
Experimental Setup



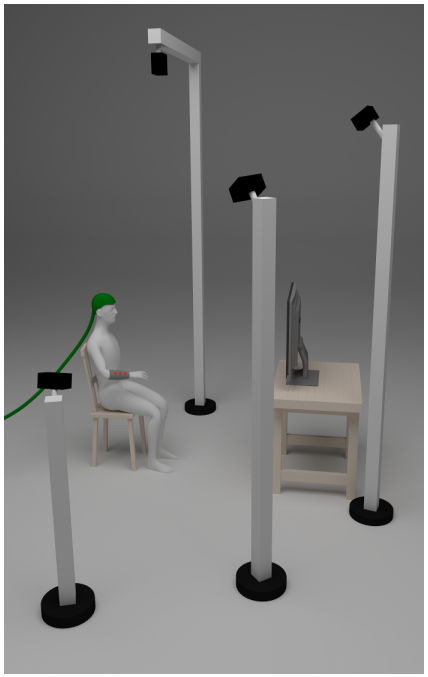


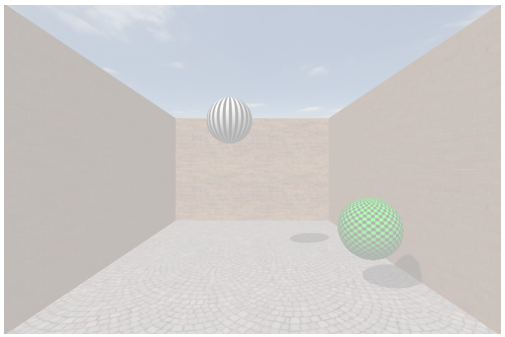


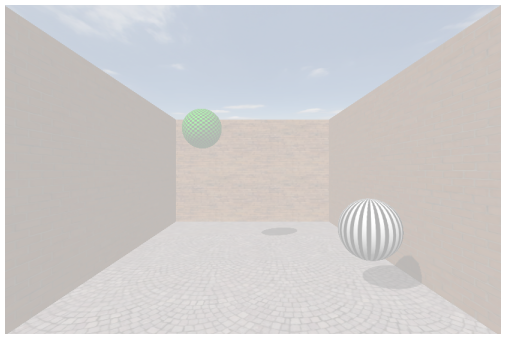


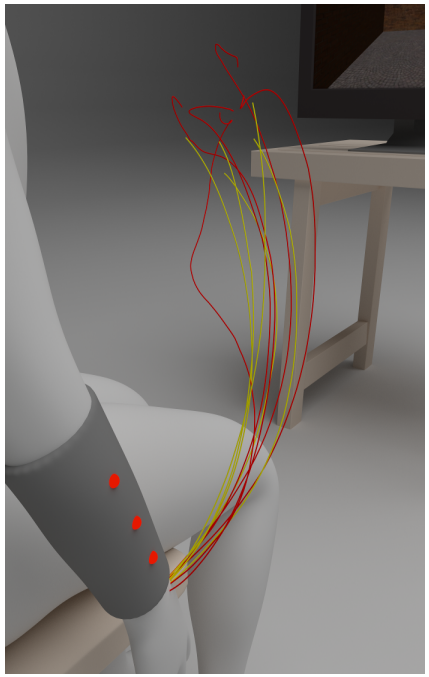


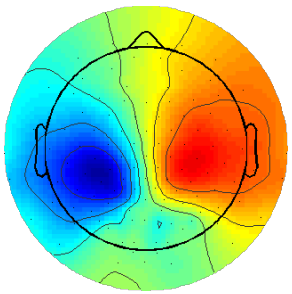
Experimental Setup

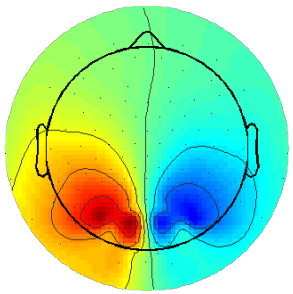


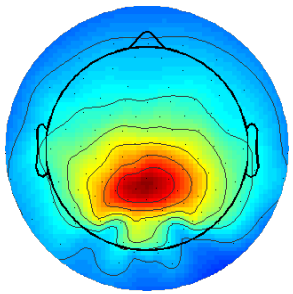


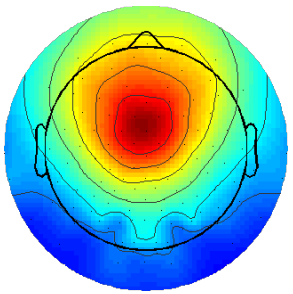


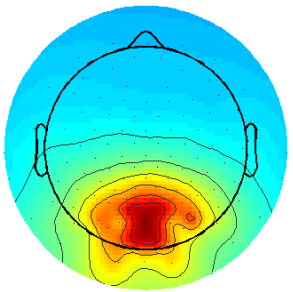


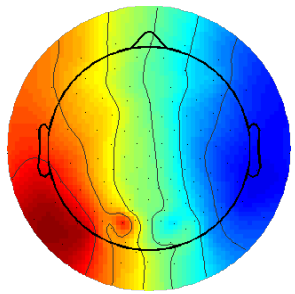




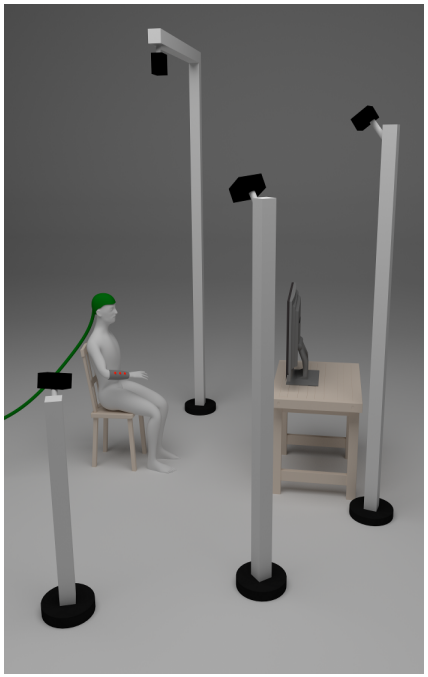


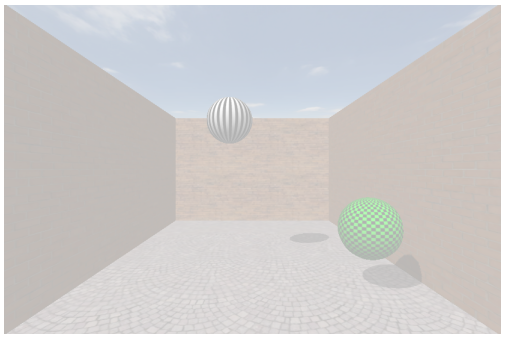


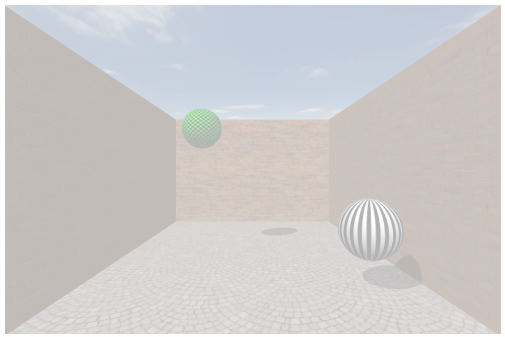


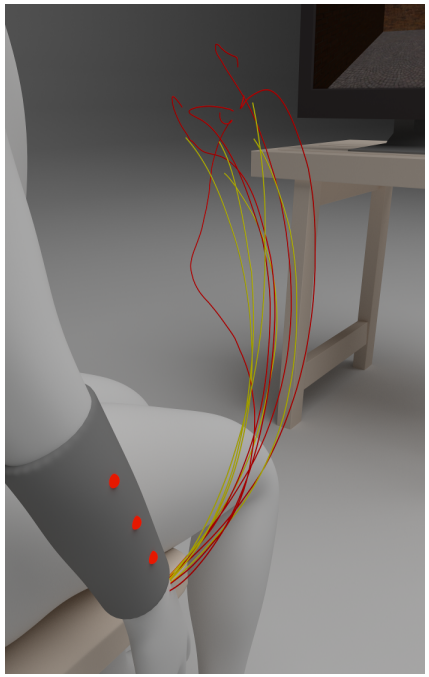


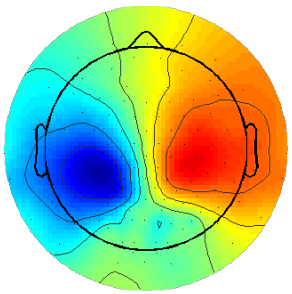
Experimental Setup

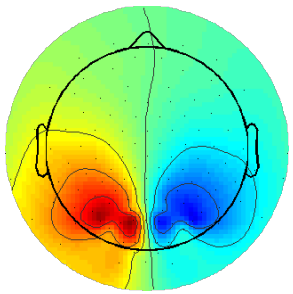


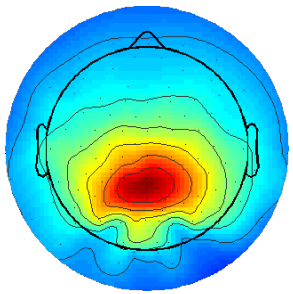


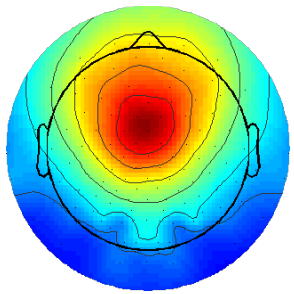


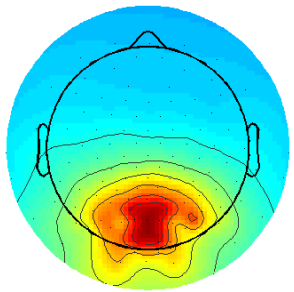


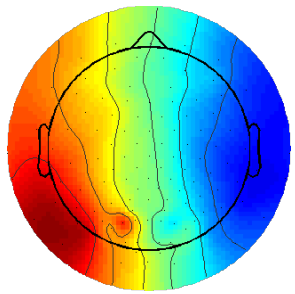












$$\{S, X_1, \ldots, X_6\} = \{\text{rest/plan}, |\alpha_{\text{IC}_1}|, \ldots, |\alpha_{\text{IC}_6}|\}$$

Data Analysis & Experimental Results

Data Analysis & Experimental Results

- Experimental data: 17 subjects with 444 - 498 trials each

Data Analysis & Experimental Results

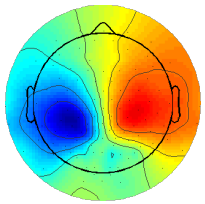
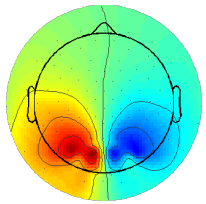
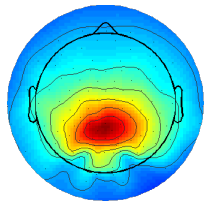
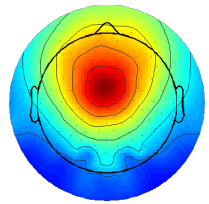
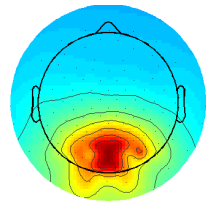
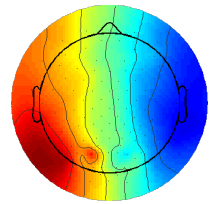
- Experimental data: 17 subjects with 444 - 498 trials each
- Encoding model: $H0: S \perp\!\!\!\perp X_i$ [HSIC (Gretton et al., *NIPS*, 2008)]

Data Analysis & Experimental Results

- Experimental data: 17 subjects with 444 - 498 trials each
- Encoding model: $H0: S \perp\!\!\!\perp X_i$ [HSIC (Gretton et al., *NIPS*, 2008)]
- Decoding model: $H0: S \perp\!\!\!\perp X_i | \mathbf{X} \backslash X_i$ [Random forest (Breiman, *Machine Learning*, 2001)]

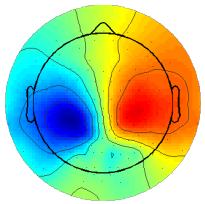
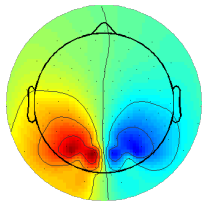
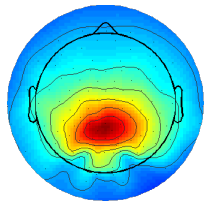
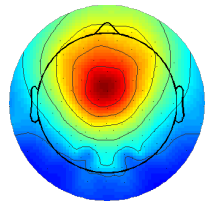
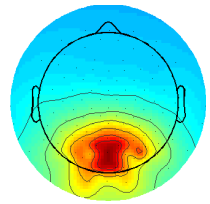
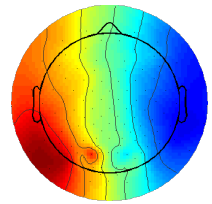
Data Analysis & Experimental Results

- Experimental data: 17 subjects with 444 - 498 trials each
- Encoding model: $H0: S \perp\!\!\!\perp X_i$ [HSIC (Gretton et al., *NIPS*, 2008)]
- Decoding model: $H0: S \perp\!\!\!\perp X_i | \mathbf{X} \backslash X_i$ [Random forest (Breiman, *Machine Learning*, 2001)]

p-values						
Encoding	0	0	0	0	0	0
Decoding	0	0	0.50	0.34	0.79	0.13

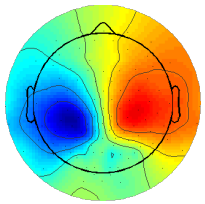
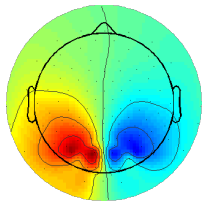
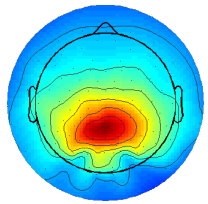
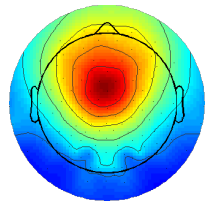
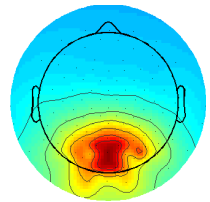
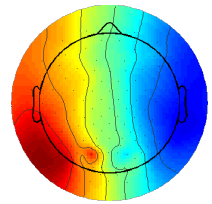
Data Analysis & Experimental Results

- Experimental data: 17 subjects with 444 - 498 trials each
- Encoding model: $H0: S \perp\!\!\!\perp X_i$ [HSIC (Gretton et al., *NIPS*, 2008)]
- Decoding model: $H0: S \perp\!\!\!\perp X_i | \mathbf{X} \backslash X_i$ [Random forest (Breiman, *Machine Learning*, 2001)]

p-values						
Encoding	0	0	0	0	0	0
Decoding	0	0	0.50	0.34	0.79	0.13

Data Analysis & Experimental Results

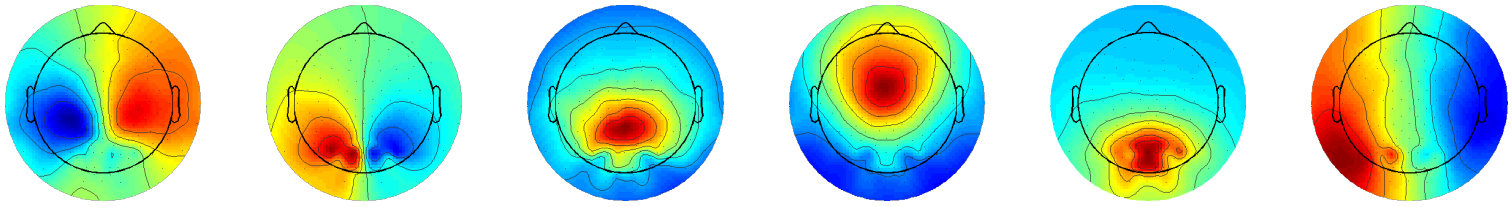
- Experimental data: 17 subjects with 444 - 498 trials each
- Encoding model: $H0: S \perp\!\!\!\perp X_i$ [HSIC (Gretton et al., *NIPS*, 2008)]
- Decoding model: $H0: S \perp\!\!\!\perp X_i | \mathbf{X} \backslash X_i$ [Random forest (Breiman, *Machine Learning*, 2001)]

p-values						
Encoding	0	0	0	0	0	0
Decoding	0	0	0.50	0.34	0.79	0.13

→ The instruction to plan a reaching movement causes modulation of α-power at every IC: $S \rightarrow \{|\alpha_{\mathrm{IC}_i}|\},\ i = 1, \ldots, 6$

Data Analysis & Experimental Results

- Experimental data: 17 subjects with 444 - 498 trials each
- Encoding model: $H0: S \perp\!\!\!\perp X_i$ [HSIC (Gretton et al., *NIPS*, 2008)]
- Decoding model: $H0: S \perp\!\!\!\perp X_i | \mathbf{X} \backslash X_i$ [Random forest (Breiman, *Machine Learning*, 2001)]

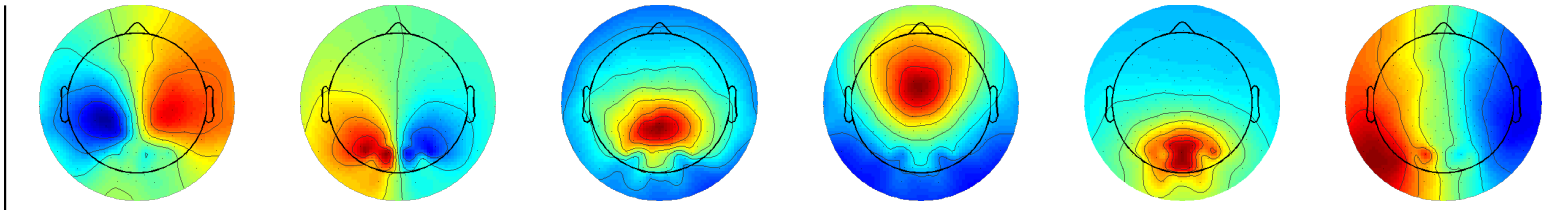


p-values						
Encoding	0	0	0	0	0	0
Decoding	0	0	0.50	0.34	0.79	0.13

$\rightarrow$ The instruction to plan a reaching movement causes modulation of α-power at every IC: $S \rightarrow \{|\alpha_{\text{IC}_i}|\},\ i = 1, \ldots, 6$

Data Analysis & Experimental Results

- Experimental data: 17 subjects with 444 - 498 trials each
- Encoding model: $H0: S \perp\!\!\!\perp X_i$ [HSIC (Gretton et al., *NIPS*, 2008)]
- Decoding model: $H0: S \perp\!\!\!\perp X_i | \mathbf{X} \backslash X_i$ [Random forest (Breiman, *Machine Learning*, 2001)]



p-values						
Encoding	0	0	0	0	0	0
Decoding	0	0	0.50	0.34	0.79	0.13

→ The instruction to plan a reaching movement causes modulation of α-power at every IC: $S \rightarrow \{|\alpha_{\mathrm{IC}_i}|\},\ i = 1, \ldots, 6$

→ Modulation of α-power at ICs 3–6 is only an indirect effect relative to ICs 1 & 2: $S \rightarrow \{|\alpha_{\mathrm{IC}_1}|, |\alpha_{\mathrm{IC}_2}|\} \rightarrow \{|\alpha_{\mathrm{IC}_3}|, |\alpha_{\mathrm{IC}_4}|, |\alpha_{\mathrm{IC}_5}|, |\alpha_{\mathrm{IC}_6}|\}$