



# CAUSAL INTERPRETATION RULES FOR ENCODING AND DECODING MODELS IN NEUROIMAGING

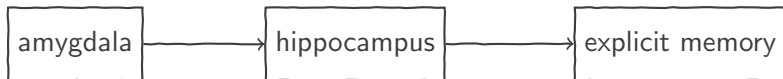
Sebastian Weichwald, Timm Meyer, Ozan Özdenizci<sup>§</sup>,  
Bernhard Schölkopf, Tonio Ball<sup>‡</sup>, Moritz Grosse-Wentrup  
MPI for Intelligent Systems, <sup>§</sup>Sabanci University, <sup>‡</sup>University of Freiburg

# Motivation



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

*(Myers et al., *Journal of Neuroscience*, 2014)*



*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.*

*(S. Hamann, Trends in Cognitive Sciences, 2001)*

1. Motivation
2. Approach
3. Encoding and decoding models in neuroimaging
4. Causal Bayesian Networks
5. Causal interpretation of encoding and decoding models
6. Empirical example
7. Wrap-up

# Approach



I'm interested in how neural activity gives rise to cognition.





I'm interested in how neural activity gives rise to cognition.

That sounds intriguing! So, what do you do?



I'm interested in how neural activity gives rise to cognition.

That sounds intriguing! So, what do you do?

I present stimuli to subjects or observe their behaviour while recording their brain activity. ...[explains common analysis methods]...



I'm interested in how neural activity gives rise to cognition.

That sounds intriguing! So, what do you do?

I present stimuli to subjects or observe their behaviour while recording their brain activity. ...[explains common analysis methods]...

Ha, interesting! ...[rephrases what has been said in causal inference slang]...



I'm interested in how neural activity gives rise to cognition.

That sounds intriguing! So, what do you do?

I present stimuli to subjects or observe their behaviour while recording their brain activity. ...[explains common analysis methods]...

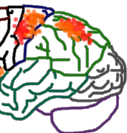
Ha, interesting! ...[rephrases what has been said in causal inference slang]...

Yeah, a solid basis for our interpretations! It also clarifies problems that we recently discussed in the community.



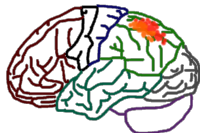
# Encoding and decoding models in neuroimaging

Trial 3



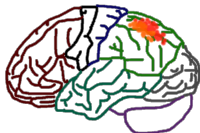
L

Trial 4



L

Trial 5



R

Trial 6



R

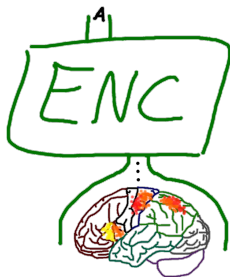
Trial 7



R



encoding



e.g. mean difference  
between conditions

decoding



e.g. classifier for  
experimental conditions





“Significant variation across experimental conditions?”

“Does removal impair decoding performance?”







“Significant variation across experimental conditions?”

“Does removal impair decoding performance?”



relevant feature  $\overset{?}{\longleftrightarrow}$  cognitive process



# Causal Bayesian Networks

- ▶  $X \rightarrow Y \iff p(Y) \neq p(Y|do\{X = x\})$
- ▶ Causal Markov Condition: d-separation  $\leadsto$  independence
- ▶ Faithfulness: d-separation  $\Leftarrow$  independence

chain

$$X \rightarrow Y \rightarrow Z$$

fork

$$X \leftarrow Y \rightarrow Z$$

collider

$$X \rightarrow Y \leftarrow Z$$



- ▶  $X \rightarrow Y \iff p(Y) \neq p(Y|do\{X = x\})$
- ▶ Causal Markov Condition: d-separation  $\leadsto$  independence
- ▶ Faithfulness: d-separation  $\Leftarrow$  independence

chain  
 $X \rightarrow Y \rightarrow Z$   
 $X \not\perp Z$

fork  
 $X \leftarrow Y \rightarrow Z$

collider  
 $X \rightarrow Y \leftarrow Z$



- ▶  $X \rightarrow Y \iff p(Y) \neq p(Y|do\{X = x\})$
- ▶ Causal Markov Condition: d-separation  $\leadsto$  independence
- ▶ Faithfulness: d-separation  $\Leftarrow$  independence

chain  
 $X \rightarrow Y \rightarrow Z$

$X \not\perp Z$   
 $X \perp Z|Y$

fork  
 $X \leftarrow Y \rightarrow Z$

collider  
 $X \rightarrow Y \leftarrow Z$



- ▶  $X \rightarrow Y \iff p(Y) \neq p(Y|do\{X = x\})$
- ▶ Causal Markov Condition: d-separation  $\leadsto$  independence
- ▶ Faithfulness: d-separation  $\Leftarrow$  independence

chain

$$X \rightarrow Y \rightarrow Z$$

$$X \not\perp Z$$

$$X \perp Z|Y$$

fork

$$X \leftarrow Y \rightarrow Z$$

$$X \not\perp Z$$

collider

$$X \rightarrow Y \leftarrow Z$$



- ▶  $X \rightarrow Y \iff p(Y) \neq p(Y|do\{X = x\})$
- ▶ Causal Markov Condition: d-separation  $\leadsto$  independence
- ▶ Faithfulness: d-separation  $\Leftarrow$  independence

chain

$$X \rightarrow Y \rightarrow Z$$

$$X \not\perp Z$$

$$X \perp Z|Y$$

fork

$$X \leftarrow Y \rightarrow Z$$

$$X \not\perp Z$$

$$X \perp Z|Y$$

collider

$$X \rightarrow Y \leftarrow Z$$



- ▶  $X \rightarrow Y \iff p(Y) \neq p(Y|do\{X = x\})$
- ▶ Causal Markov Condition: d-separation  $\leadsto$  independence
- ▶ Faithfulness: d-separation  $\Leftarrow$  independence

chain

$$X \rightarrow Y \rightarrow Z$$

$$X \not\perp Z$$

$$X \perp Z|Y$$

fork

$$X \leftarrow Y \rightarrow Z$$

$$X \not\perp Z$$

$$X \perp Z|Y$$

collider

$$X \rightarrow Y \leftarrow Z$$

$$X \perp Z$$





- ▶  $X \rightarrow Y \iff p(Y) \neq p(Y|do\{X = x\})$
- ▶ Causal Markov Condition: d-separation  $\leadsto$  independence
- ▶ Faithfulness: d-separation  $\Leftarrow$  independence

chain

$$X \rightarrow Y \rightarrow Z$$

$$X \not\perp Z$$

$$X \perp Z|Y$$

fork

$$X \leftarrow Y \rightarrow Z$$

$$X \not\perp Z$$

$$X \perp Z|Y$$

collider

$$X \rightarrow Y \leftarrow Z$$

$$X \perp Z$$

$$X \not\perp Z|Y$$





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

(Myers et al., *Journal of Neuroscience*, 2014)



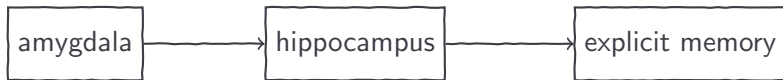


*We tested [...] whether pre-stimulus alpha oscillations measured with electroencephalography (EEG) **influence** the encoding of items into working memory.*

(Myers et al., *Journal of Neuroscience*, 2014)

$\alpha \not\perp \text{WM}$

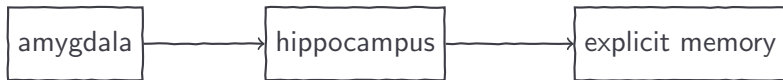




*Hippocampal activity in this study was correlated with amygdala activity, supporting the view that the amygdala **enhances** explicit memory by **modulating** activity in the hippocampus.*

(S. Hamann, *Trends in Cognitive Sciences*, 2001)





*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.*

(S. Hamann, *Trends in Cognitive Sciences*, 2001)

AM  $\not\perp$  EM

AM  $\perp$  EM | HC



# Causal interpretation of encoding and decoding models

Let's set out the causal component of already performed analyses..



Let's set out the causal component of already performed analyses..

stimulus- vs response-based

feature relevance  $\leftrightarrow$  marginal/conditional dependence

$\leadsto$  16 causal interpretation rules





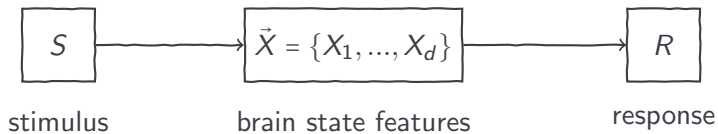
Let's set out the causal component of already performed analyses..

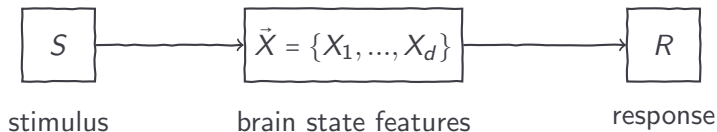
stimulus- vs response-based

feature relevance  $\leftrightarrow$  marginal/conditional dependence

$\leadsto$  16 causal interpretation rules  
simple

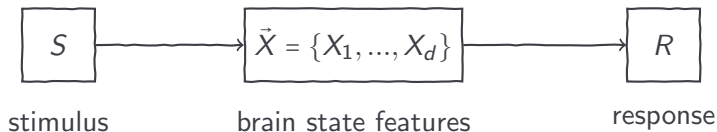






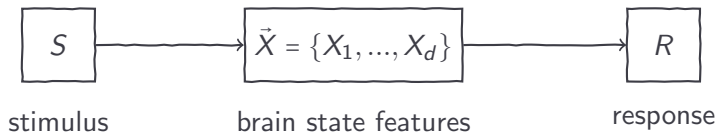
stimulus-based		response-based	
$p(\vec{X} S)$	encoding decoding	$p(\vec{X} R)$	
$p(S \vec{X})$		$p(R \vec{X})$	





stimulus-based			response-based	
$p(\vec{X} S)$	causal	encoding		$p(\vec{X} R)$
$p(S \vec{X})$		decoding	causal	$p(R \vec{X})$





stimulus-based			response-based	
$p(\vec{X} S)$	causal	encoding	<i>anti</i> -causal	$p(\vec{X} R)$
$p(S \vec{X})$	<i>anti</i> -causal	decoding	causal	$p(R \vec{X})$







$$p(X_i|C = c_1) \stackrel{?}{\neq} p(X_i|C = c_2)$$





$$p(X_i|C = c_1) \stackrel{?}{\neq} p(X_i|C = c_2)$$

$$X_i \not\perp C$$







$$p(X_i|C = c_1) \stackrel{?}{\neq} p(X_i|C = c_2)$$

$$X_i \not\perp C$$



$$p(C|\vec{X}) \stackrel{?}{\neq} p(C|\vec{X} \setminus X_i)$$





$$p(X_i|C = c_1) \stackrel{?}{\neq} p(X_i|C = c_2)$$

$$X_i \not\perp C$$



$$p(C|\vec{X}) \stackrel{?}{\neq} p(C|\vec{X} \setminus X_i)$$

$$X_i \not\perp C|\vec{X} \setminus X_i$$





# Causal interpretation rules (1)

	Feature $X_i$ relevant?		Causal interpretation
	Encoding	Decoding	
Stimulus-based	<div>×</div> <div>✓</div>	<div>×</div> <div>✓</div>	
Response-based	<div>×</div> <div>✓</div>	<div>×</div> <div>✓</div>	



# Causal interpretation rules (1)

	Feature $X_i$ relevant?		Causal interpretation
	Encoding	Decoding	
Stimulus-based	×		no effect of $S$
	✓		
		×	
		✓	
Response-based	×		
	✓		
		×	
		✓	



# Causal interpretation rules (1)

	Feature $X_i$ relevant?		Causal interpretation
	Encoding	Decoding	
Stimulus-based	×		no effect of $S$
	✓		effect of $S$
		×	
		✓	
Response-based	×		
	✓		
		×	
		✓	



# Causal interpretation rules (1)

	Feature $X_i$ relevant?		Causal interpretation
	Encoding	Decoding	
Stimulus-based	×		no effect of $S$
	✓	$X_i \perp\!\!\!\perp S   \vec{X} \setminus X_i$ $S \rightarrow X_j \rightarrow X_i$	effect of $S$
Response-based	×	$S \rightarrow X_i$	
	✓		
		×	
		✓	



# Causal interpretation rules (1)

	Feature $X_i$ relevant?		Causal interpretation
	Encoding	Decoding	
Stimulus-based	×		no effect of $S$
	✓		effect of $S$
		×	
		✓	
Response-based	×		
	✓		
		×	
		✓	





# Causal interpretation rules (1)

	Feature $X_i$ relevant?		Causal interpretation
	Encoding	Decoding	
Stimulus-based	×		no effect of $S$
	✓		effect of $S$
		×	inconclusive
		✓	
Response-based	×		
	✓		
		×	
		✓	



# Causal interpretation rules (1)

	Feature $X_i$ relevant?		Causal interpretation
	Encoding	Decoding	
Stimulus-based	×		no effect of $S$
	✓		effect of $S$
		$X_i \not\perp S   \vec{X} \setminus X_i$ $S \rightarrow X_i$	inconclusive
Response-based	×	$S \rightarrow X_j \leftarrow X_i$	
	✓		
		$\times$ $\checkmark$	



# Causal interpretation rules (1)

	Feature $X_i$ relevant?		Causal interpretation
	Encoding	Decoding	
Stimulus-based	×		no effect of $S$
	✓		effect of $S$
		×	inconclusive
		✓	
Response-based	×		
	✓		
		×	
		✓	



# Causal interpretation rules (1)

	Feature $X_i$ relevant?		Causal interpretation
	Encoding	Decoding	
Stimulus-based	×		no effect of $S$
	✓		effect of $S$
		×	inconclusive
		✓	inconclusive
Response-based	×		
	✓		
		×	
		✓	



# Causal interpretation rules (1)

	Feature $X_i$ relevant?		Causal interpretation
	Encoding	Decoding	
Stimulus-based	×		no effect of $S$
	✓		effect of $S$
		×	inconclusive
		✓	inconclusive
Response-based	×		no cause of $R$
	✓		
		×	
		✓	



# Causal interpretation rules (1)

	Feature $X_i$ relevant?		Causal interpretation
	Encoding	Decoding	
Stimulus-based	×		no effect of $S$
	✓	$X_i \not\perp R$	effect of $S$
		$X_i \leftarrow h \rightarrow R$	inconclusive
		✓ $X_i \rightarrow R$	inconclusive
Response-based	×		no cause of $R$
	✓	example (A) $\alpha \rightarrow WM$	
		×	
		✓	



# Causal interpretation rules (1)

	Feature $X_i$ relevant?		Causal interpretation
	Encoding	Decoding	
Stimulus-based	×		no effect of $S$
	✓		effect of $S$
		×	inconclusive
		✓	inconclusive
Response-based	×		no cause of $R$
	✓		
		×	
		✓	



# Causal interpretation rules (1)

	Feature $X_i$ relevant?		Causal interpretation
	Encoding	Decoding	
Stimulus-based	×		no effect of $S$
	✓		effect of $S$
		×	inconclusive
		✓	inconclusive
Response-based	×		no cause of $R$
	✓		inconclusive
		×	
		✓	





# Causal interpretation rules (1)

	Feature $X_i$ relevant?		Causal interpretation
	Encoding	Decoding	
Stimulus-based	×		no effect of $S$
	✓		effect of $S$
		$X_i \perp\!\!\!\perp R   \vec{X} \setminus X_i$	inconclusive
		$X_i \rightarrow X_j \rightarrow R$	inconclusive
Response-based	×		no cause of $R$
	✓		inconclusive
		$X_i$	
		$R$	
		×	
		✓	



# Causal interpretation rules (1)

	Feature $X_i$ relevant?		Causal interpretation
	Encoding	Decoding	
Stimulus-based	×		no effect of $S$
	✓		effect of $S$
		×	inconclusive
		✓	inconclusive
Response-based	×		no cause of $R$
	✓		inconclusive
		×	
		✓	



# Causal interpretation rules (1)

	Feature $X_i$ relevant?		Causal interpretation
	Encoding	Decoding	
Stimulus-based	×		no effect of $S$
	✓		effect of $S$
		×	inconclusive
		✓	inconclusive
Response-based	×		no cause of $R$
	✓		inconclusive
		×	inconclusive
		✓	



# Causal interpretation rules (1)

	Feature $X_i$ relevant?		Causal interpretation
	Encoding	Decoding	
Stimulus-based	×		no effect of $S$
	✓	$X_i \not\perp R   \vec{X} \setminus X_i$	effect of $S$
		$X_i \rightarrow R$	inconclusive
		✓	inconclusive
Response-based		$X_i \leftarrow h \rightarrow R$	
	×	↓	no cause of $R$
	✓	$X_j$	inconclusive
		×	inconclusive
		✓	



# Causal interpretation rules (1)

	Feature $X_i$ relevant?		Causal interpretation
	Encoding	Decoding	
Stimulus-based	×		no effect of $S$
	✓		effect of $S$
		×	inconclusive
		✓	inconclusive
Response-based	×		no cause of $R$
	✓		inconclusive
		×	inconclusive
		✓	



# Causal interpretation rules (1)

	Feature $X_i$ relevant?		Causal interpretation
	Encoding	Decoding	
Stimulus-based	×		no effect of $S$
	✓		effect of $S$
		×	inconclusive
		✓	inconclusive
Response-based	×		no cause of $R$
	✓		inconclusive
		×	inconclusive
		✓	inconclusive



	Feature $X_i$ relevant?		Causal interpretation
	Encoding	Decoding	
Stimulus-based	×	×	no effect of $S$
	✓	×	effect of $S$
	×	✓	inconclusive
	✓	✓	inconclusive
Are decoding models useful at all?			
Response-based	×	×	no cause of $R$
	✓	×	inconclusive
	×	✓	inconclusive
	✓	✓	inconclusive



## Causal interpretation rules (2)

	Feature $X_i$ relevant?		Causal interpretation
	Encoding	Decoding	
Stimulus-based	×	×	
	✓	×	
	×	✓	
	✓	✓	
Response-based	×	×	
	✓	×	
	×	✓	
	✓	✓	





# Causal interpretation rules (2)

	Feature $X_i$ relevant?		Causal interpretation
	Encoding	Decoding	
Stimulus-based	×	×	
	✓	×	
	×	✓	
	✓	✓	
Response-based	×	×	
	✓	×	
	×	✓	
	✓	✓	



# Causal interpretation rules (2)

	Feature $X_i$ relevant?		Causal interpretation
	Encoding	Decoding	
Stimulus-based	×	×	no effect of $S$
	✓	×	
	×	✓	provides context
	✓	✓	effect of $S$
Response-based	×	×	no cause of $R$
	✓	×	
	×	✓	provides context
	✓	✓	inconclusive



	Feature $X_i$ relevant?		Causal interpretation
	Encoding	Decoding	
Stimulus-based	x	x	no effect of $S$
	✓	x	provides context
	x	✓	
	✓	✓	effect of $S$
Response-based	x	e.g. $S \rightarrow X_j \rightarrow X_i$	cause of $R$
	✓	x	provides context
	x	✓	
	✓	✓	inconclusive

$$X_i \not\perp S \text{ and } X_i \perp S | \vec{X} \setminus X_i$$

$S \rightarrow X_i$  indirectly



# Causal interpretation rules (2)

	Feature $X_i$ relevant?		Causal interpretation
	Encoding	Decoding	
Stimulus-based	×	×	no effect of $S$
	✓	×	
	×	✓	provides context
	✓	✓	effect of $S$
Response-based	×	×	no cause of $R$
	✓	×	
	×	✓	provides context
	✓	✓	inconclusive



	Feature $X_i$ relevant?		Causal interpretation
	Encoding	Decoding	
Stimulus-based	×	×	no effect of $S$
	✓	×	indirect effect of $S$
	×	✓	provides context
	✓	✓	effect of $S$
Response-based	×	×	no cause of $R$
	✓	×	
	×	✓	provides context
	✓	✓	inconclusive



# Causal interpretation rules (2)

	Feature $X_i$ relevant?		Causal interpretation
	Encoding	Decoding	
Stimulus-based	x	x	no effect of $S$
	✓	x	direct effect of $S$
	x	✓	provides context
	✓	✓	effect of $S$
Response-based	x	x	no cause of $R$
	✓	x	example (B) $AM \rightarrow HC \rightarrow EM$
	x	✓	provides context
	✓	✓	inconclusive



	Feature $X_i$ relevant?		Causal interpretation
	Encoding	Decoding	
Stimulus-based	×	×	no effect of $S$
	✓	×	indirect effect of $S$
	×	✓	provides context
	✓	✓	effect of $S$
Response-based	×	×	no cause of $R$
	✓	×	
	×	✓	provides context
	✓	✓	inconclusive

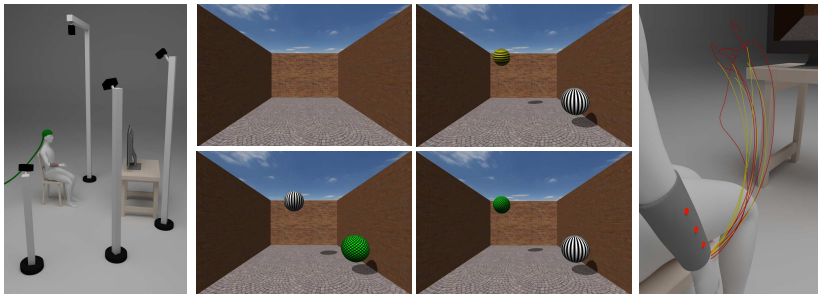


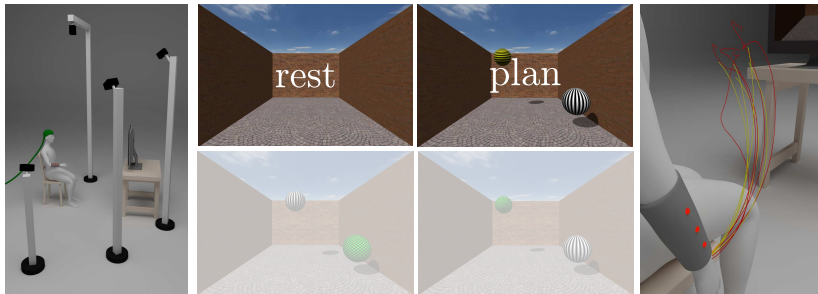
	Feature $X_i$ relevant?		Causal interpretation
	Encoding	Decoding	
Stimulus-based	×	×	no effect of $S$
	✓	×	indirect effect of $S$
	×	✓	provides context
	✓	✓	effect of $S$
Response-based	×	×	no cause of $R$
	✓	×	no direct cause of $R$
	×	✓	provides context
	✓	✓	inconclusive



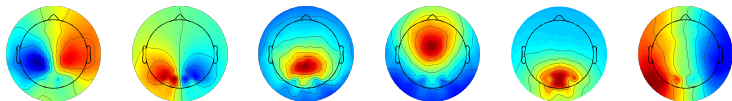


# Empirical example

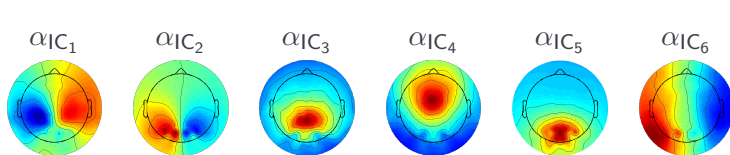




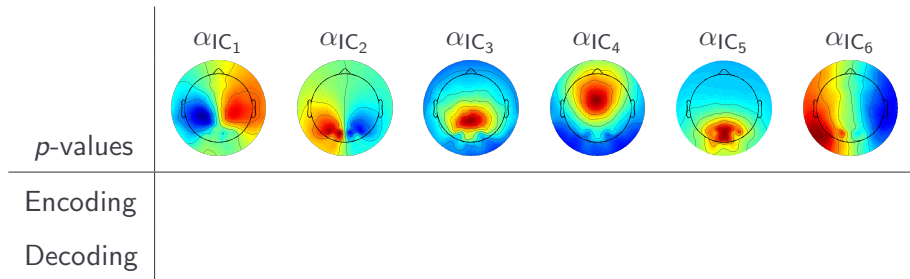
## (Relevant) Features









## (Relevant) Features



## (Relevant) Features









## (Relevant) Features

	$\alpha C_1$	$\alpha C_2$	$\alpha C_3$	$\alpha C_4$	$\alpha C_5$	$\alpha C_6$
<i>p</i> -values						
Encoding	$10^{-3}$	$10^{-3}$	$10^{-3}$	$10^{-3}$	$10^{-3}$	$10^{-3}$
Decoding	$10^{-3}$	$10^{-3}$	0.50	0.34	0.79	0.13

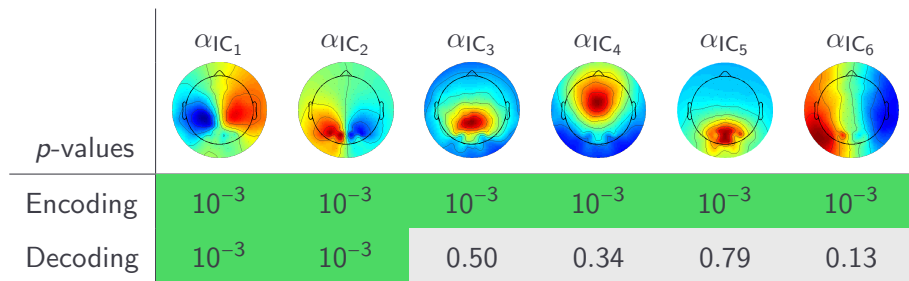


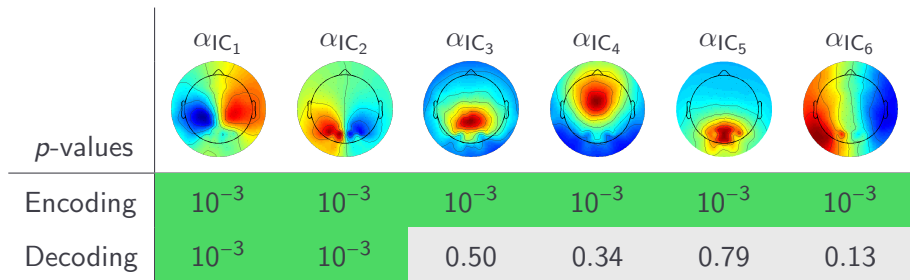
## (Relevant) Features

	$\alpha C_1$	$\alpha C_2$	$\alpha C_3$	$\alpha C_4$	$\alpha C_5$	$\alpha C_6$
<i>p</i> -values						
Encoding	$10^{-3}$	$10^{-3}$	$10^{-3}$	$10^{-3}$	$10^{-3}$	$10^{-3}$
Decoding	$10^{-3}$	$10^{-3}$	0.50	0.34	0.79	0.13



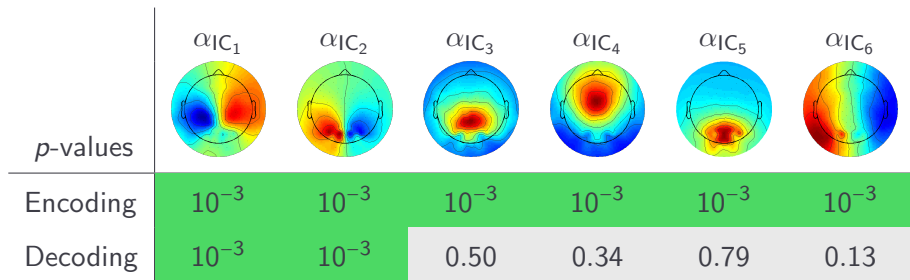




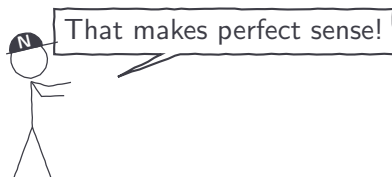


- ▶ instruction to plan a reaching movement is causal for all  $\alpha|C_i$
- ▶  $\alpha|C_3, \dots, \alpha|C_6$  are only indirect effects





- ▶ instruction to plan a reaching movement is causal for all  $\alpha_{IC_i}$
- ▶  $\alpha_{IC_3}, \dots, \alpha_{IC_6}$  are only indirect effects



Wrap-up



feature relevance



feature relevance  $\leftrightarrow$  (conditional) (in)dependence



feature relevance  $\leftrightarrow$  (conditional) (in)dependence  $\leftrightarrow$  causal structure





feature relevance  $\leftrightarrow$  (conditional) (in)dependence  $\leftrightarrow$  causal structure

- simple interpretation rules
- reinterpretation of previous results?
- resolve recently discussed issues



feature relevance  $\leftrightarrow$  (conditional) (in)dependence  $\leftrightarrow$  causal structure

- simple interpretation rules
- reinterpretation of previous results?
- resolve recently discussed issues

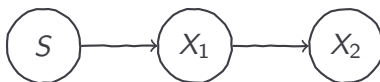
It's an interesting application!



Sebastian Weichwald, Timm Meyer, Ozan Özdenizci, Bernhard Schölkopf, Tonio Ball, Moritz Grosse-Wentrup:

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





- ▶ Establish  $X_1 \rightarrow X_2$  even in the presence of latent confounders  
(Grosse-Wentrup et al., under revision)
- ▶ Identify  $X_2$  from a linear mixture of signals

- Problem of confounds in MVPA

(Todd et al., *NeuroImage*, 2013; Woolgar et al., *NeuroImage*, 2014)

- Interpretation of weight vectors of linear models

(Haufe et al., *NeuroImage*, 2014)

- ▶ Type II errors
- ▶ Conditional independence tests
  - permutation-based  $\leadsto$  biased  
(Strobl et al., *BMC Bioinformatics*, 2008)
  - unbiased  $\leadsto$  hard  
(Zhang et al., *UAI*, 2011)
- ▶ Untestable assumption: faithfulness