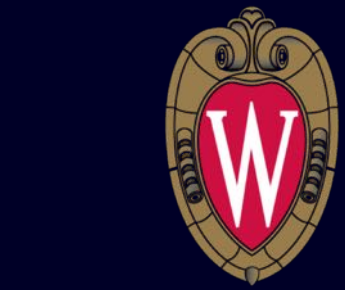



# A quantitative account of agents and their actions

using the causal framework of Integrated Information Theory (IIT)

*Larissa Albantakis*

# What is an agent?

Agents are open systems that dynamically and informationally interact with the environment.


https://ourstoriesandperspectives.com/2013/03/06/a-wonder-of-winter-a-harbinger-of-spring/

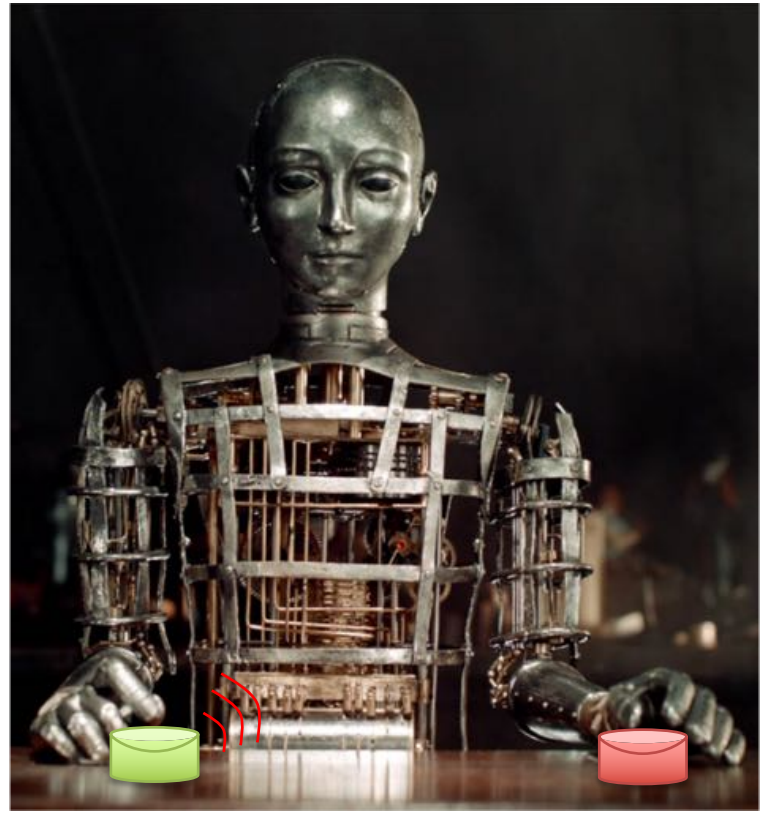
# What is an agent?

Agents are open systems that dynamically and informationally interact with the environment.

## 2. How do we distinguish autonomous actions from mere reflexes?

### (A) autonomous

### (B) automatic



From movie “Hugo” 2011.

# What is an agent?

Agents are open systems that dynamically and informationally interact with the environment.

## 2. How do we distinguish autonomous actions from mere reflexes?

**(A) autonomous**

**(B) automatic**

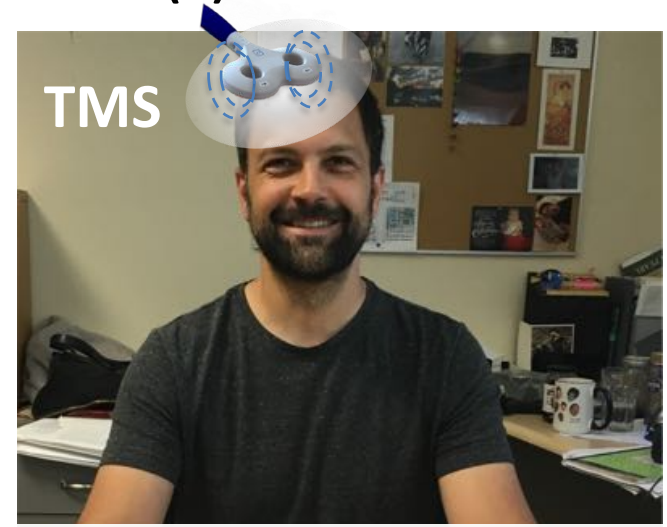
# What is an agent?

Agents are open systems that dynamically and informationally interact with the environment.

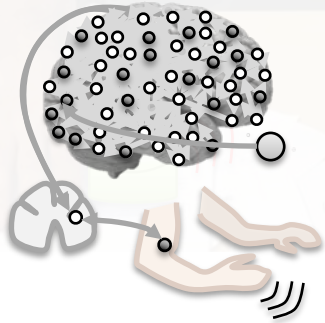
## 2. How do we distinguish autonomous actions from mere reflexes?

**(A) autonomous**

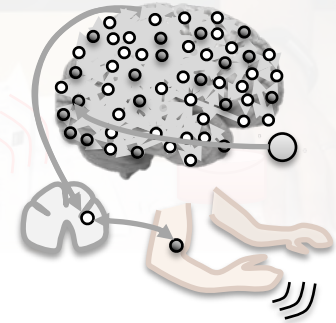
**(B) automatic**




What was the cause for clicking right?



"Neurons were activated and interacted with other neurons which then triggered a motor response."



# What is an agent?

A minimal working definition:

## An autonomous agent is

**(1) an open system with stable, self-defined and self-maintained causal borders,**

**(2) with the capacity to perform actions that are (partially) caused from within**

Can we quantify this?

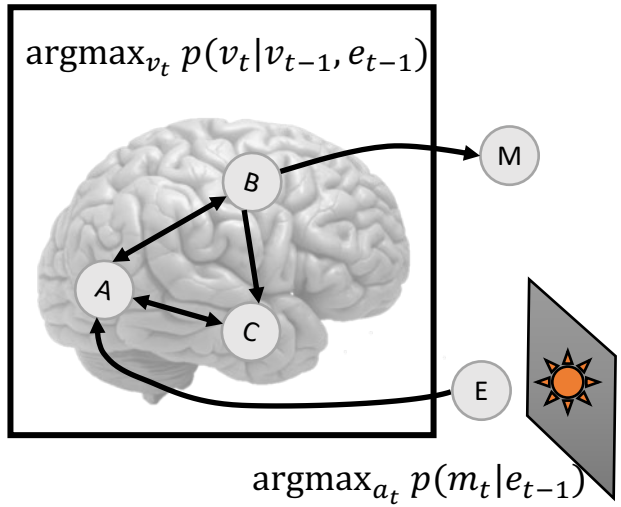
Recent related work:


- Bertschinger N, Olbrich E, Ay N, Jost J (2008) Autonomy: An information theoretic perspective. Biosystems 91:331–345.
- Krakauer D, Bertschinger N, Olbrich E, Ay N, Flack JC (2014) The Information Theory of Individuality. Arxiv 1412.2447.
- Biehl M, Ikegami T, Polani D (2016) Towards information based spatiotemporal patterns as a foundation for agent representation in dynamical systems. In: Artificial Life Conference 2016, pp 722.
- Kolchinsky A, Wolpert DH (2018) Semantic information, autonomous agency and non-equilibrium statistical physics. Interface Focus 8:20180041.
- Kirchhoff M, Parr T, Palacios E, Friston K, Kiverstein J (2018) The Markov blankets of life: autonomy, active inference and the free energy principle. J R Soc Interface 15:20170792.
- Walker SI, Davies PCW (2013) The algorithmic origins of life. J R Soc Interface 10:20120869.

# Extrinsic vs. intrinsic information

extrinsic and correlational

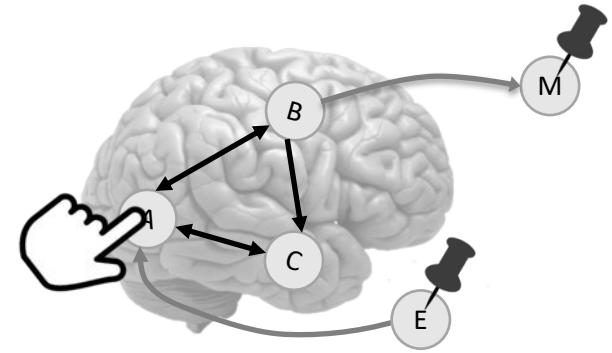
**Predictive information**

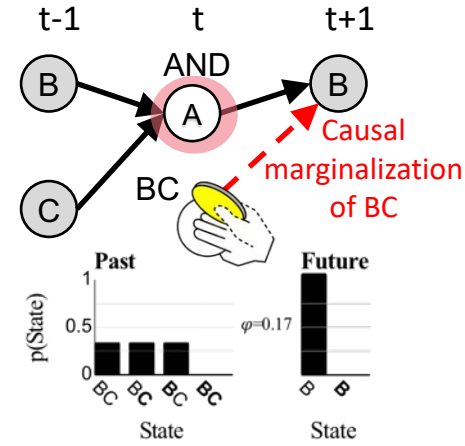



intrinsic and causal

**Intrinsic information**

$p(v_{t\pm1}|do(a_t), BC_t = ?)$



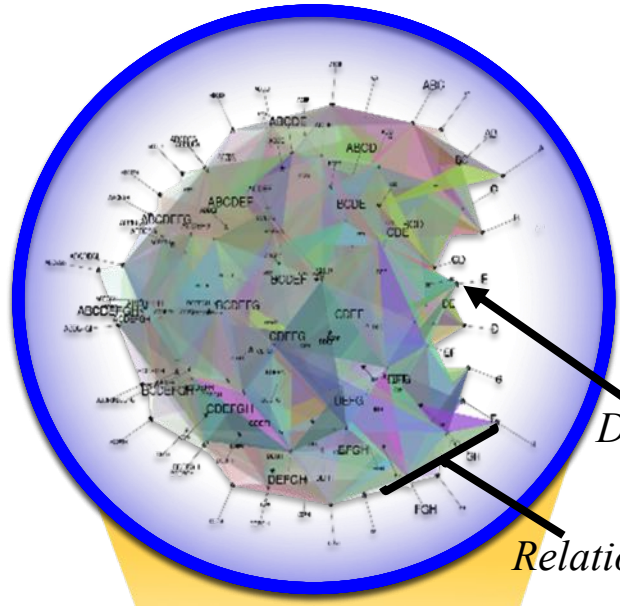



# Integrated Information Theory (IIT) of Consciousness

**An experience is a maximum (Φ) of cause-effect power (a cause-effect structure)**

**Experience**

=



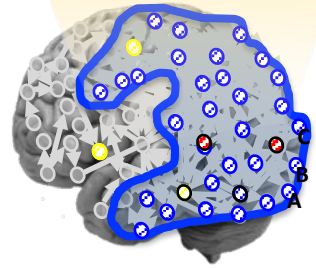
*Distinctions*

*Relations*

**Cause-effect structure**

***Quantity***:
*irreducibility Φ*
*of the cause-effect structure*

***Quality:***
*'form'*
*of the cause-effect structure*



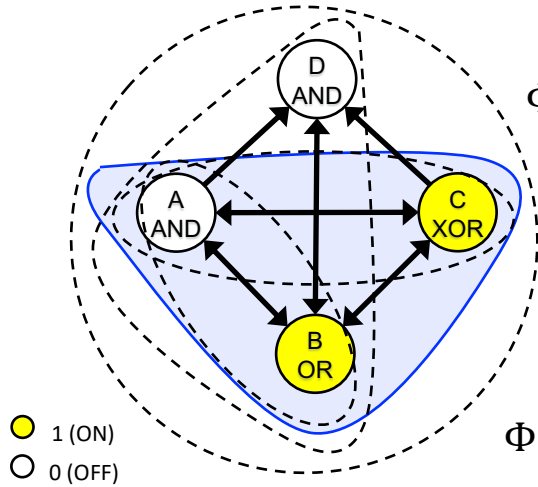
$\Phi^{max}$

**Physical substrate**

***Tononi et al***
***Nature Rev Neurosci 2016***

# Causal framework of IIT

## 1. Identifying self-defined causal borders – maxima of Φ




$\Phi(\mathrm{ABCD}) = 0.24$

$\boldsymbol{\Phi^{max} = \Phi(ABC) = 1.82}$

$\Phi(\mathrm{AC}) = 1.0$

$\Phi(\mathrm{AB}) = 1.0$

$\Phi = 0.0$ for all remaining sets

An autonomous agent is:

**1. an open system with stable, self-defined and self-maintained causal borders,**

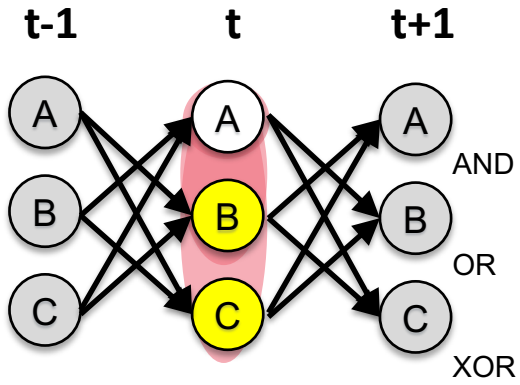
2. with the capacity to perform actions that are (partially) caused from within

Oizumi M, Albantakis L, Tononi G (2014) From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0. PLoS Comput Biol 10:e1003588.

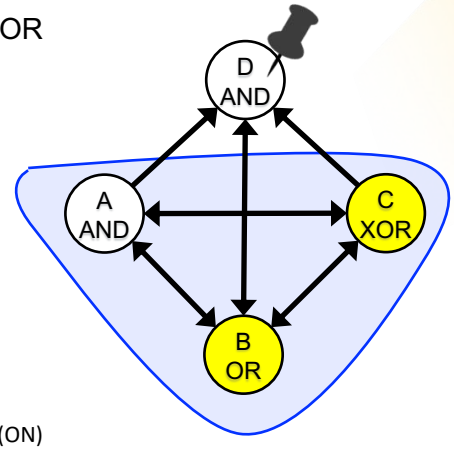
# Causal framework of IIT

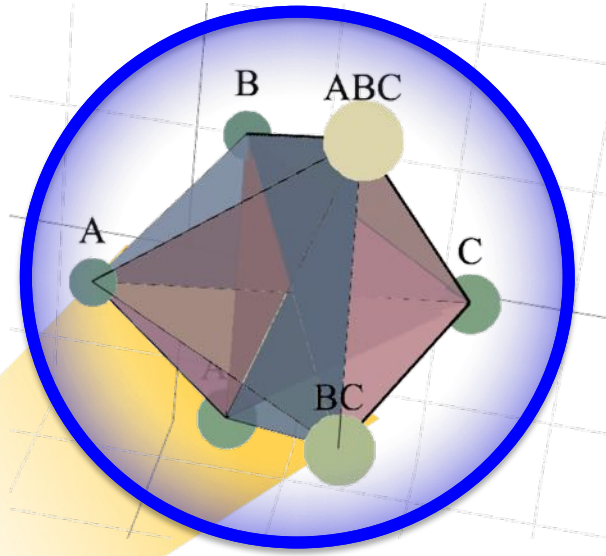
## 1. Identifying self-defined causal borders – maxima of Φ

**Unfolding the causes and effects of a system in a state**




$$\Phi(\mathbf{ABC}) = 1.82$$



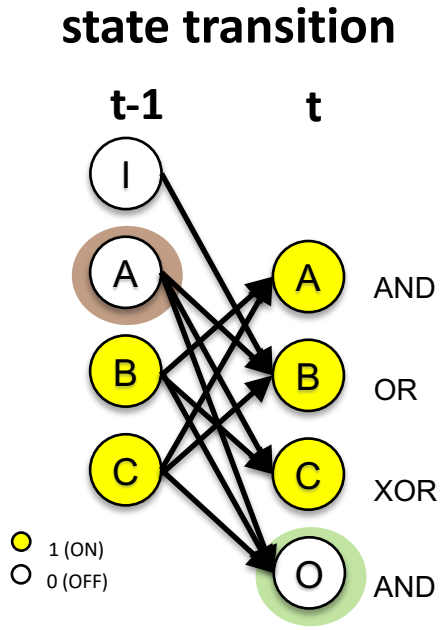




1 (ON)
0 (OFF)

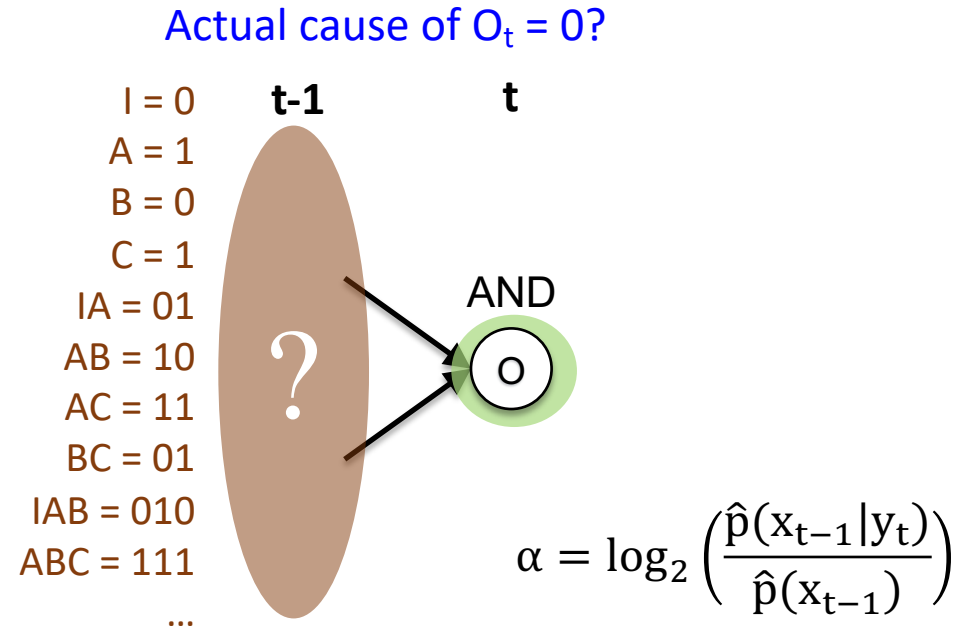
Oizumi M, Albantakis L, Tononi G (2014) From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0. PLoS Comput Biol 10:e1003588.

# Causal framework of IIT

## 2. Identifying actual causes (of actions) – maxima of $\alpha$




$\alpha^{\max} = 0.19$ bit




$$\alpha = \log_2\left(\frac{\hat{p}(x_{t-1}|y_t)}{\hat{p}(x_{t-1})}\right)$$

An autonomous agent is:
1. an open system with stable, self-defined and self-maintained causal borders,
**2. with the capacity to perform actions that are (partially) caused from within**

Albantakis L, Marshall W, Hoel E, Tononi G (2019) What caused what? A quantitative account of actual causation using dynamical causal networks. Entropy 21:459.

An autonomous agent is:
**1. an open system with stable, self-defined and self-maintained causal borders,**
2. with the capacity to perform actions that are (partially) caused from within

# 1. Identifying self-defined causal borders

Example 1: Fission yeast cell cycle model

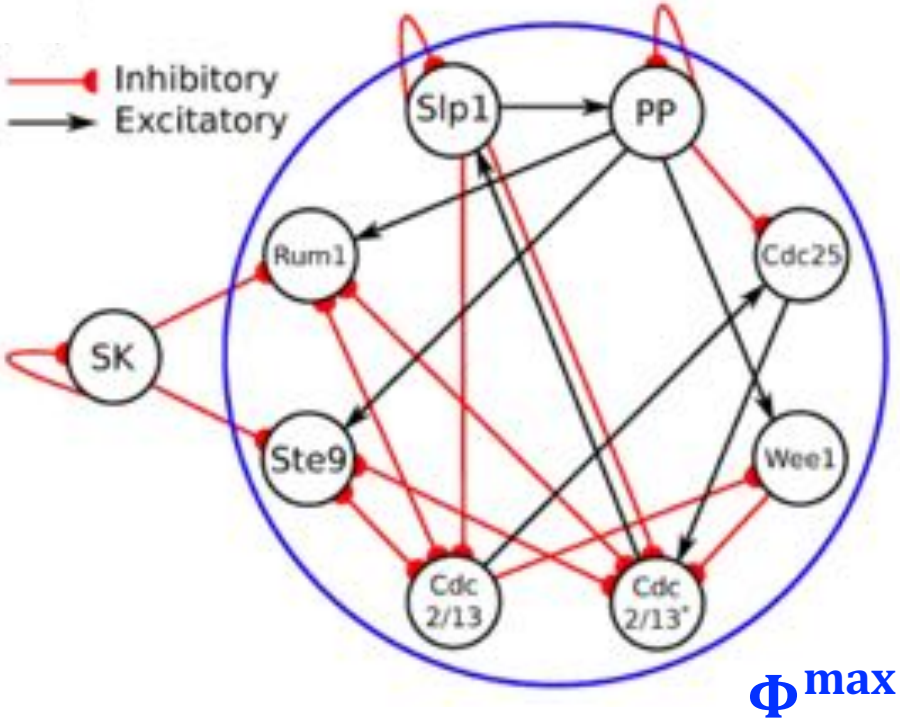
Marshall W, Kim H, Walker SI, Tononi G, Albantakis L (2017)

Example 2: Evolved artificial agents (animats)

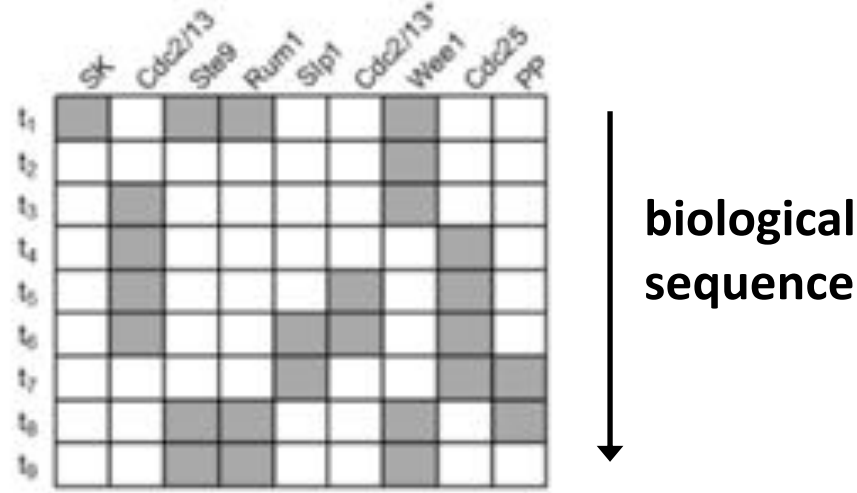
Albantakis L, Hintze A, Koch C, Adami C, Tononi G (2014)

# The yeast cell-cycle model has stable causal borders throughout its biological sequence

**Boolean network model of the fission-yeast cell-cycle**




$\Phi^{max}$

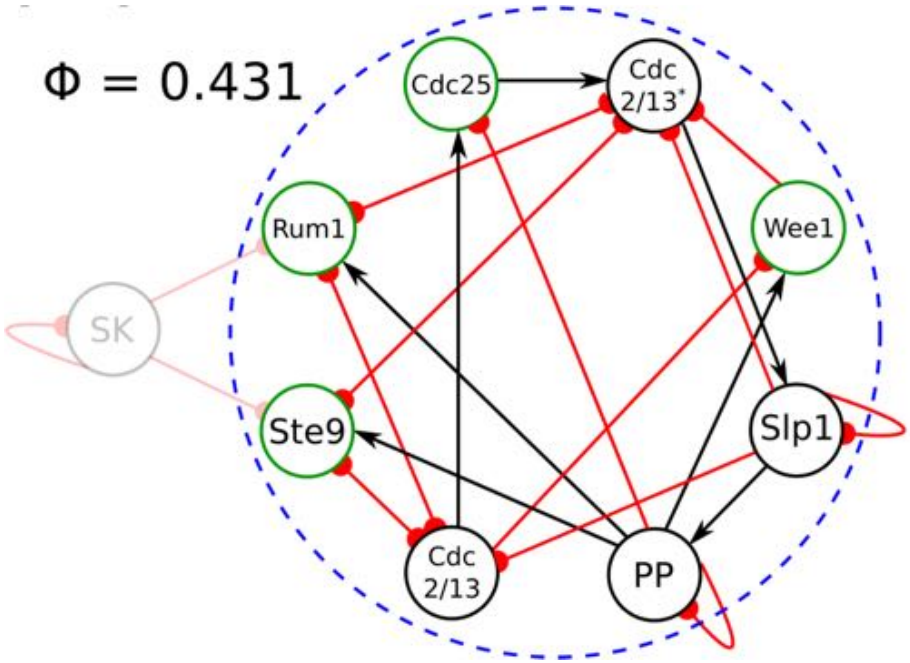


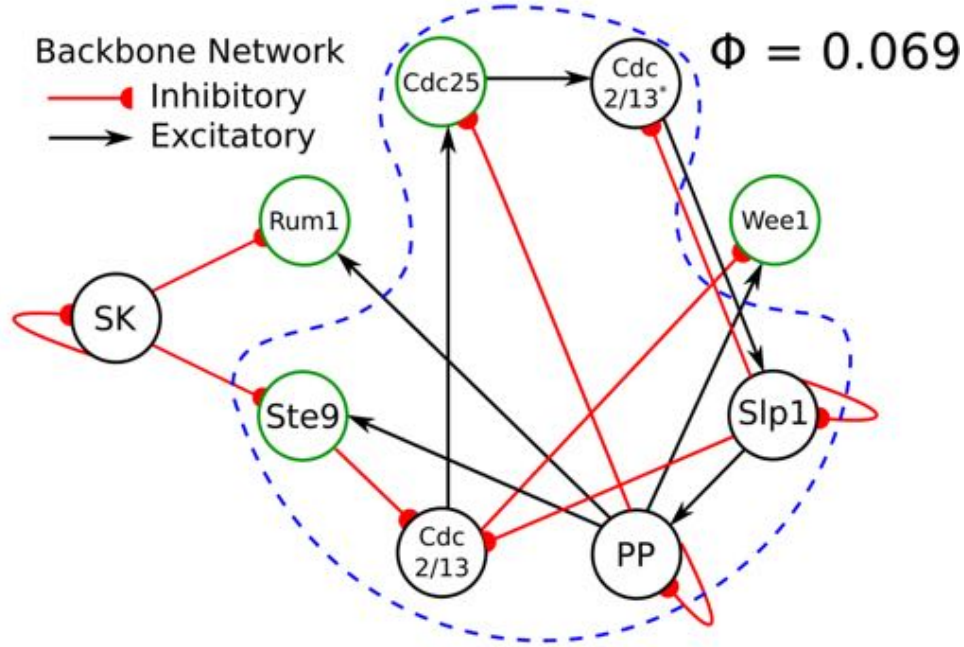

Marshall W, Albantakis L, Tononi G (2018) Black-boxing and cause-effect power. PLOS Comput Biol 14:e1006114.

# A functionally equivalent reduced version of the network does not form a stable entity

fission yeast cell cycle model

functionally identical
'backbone' network

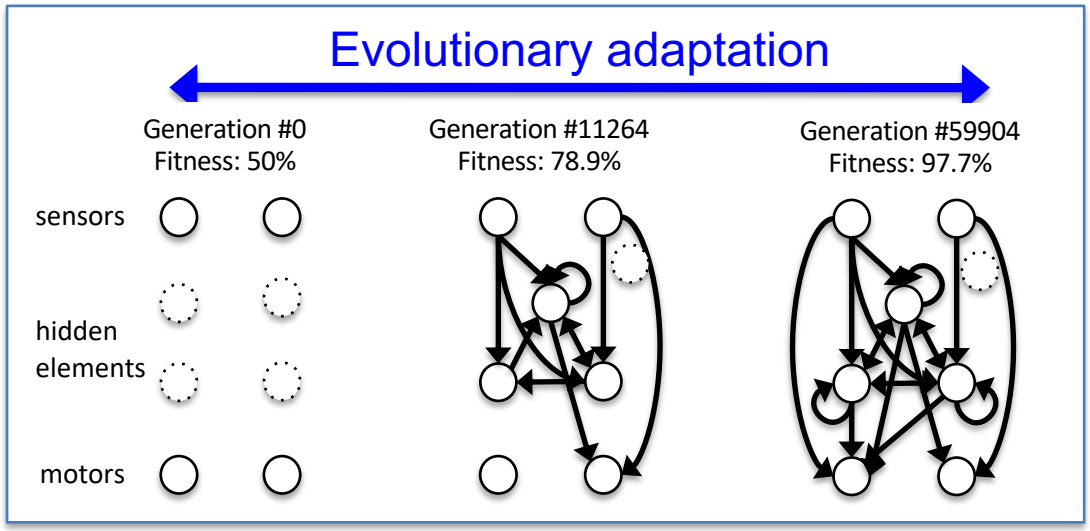


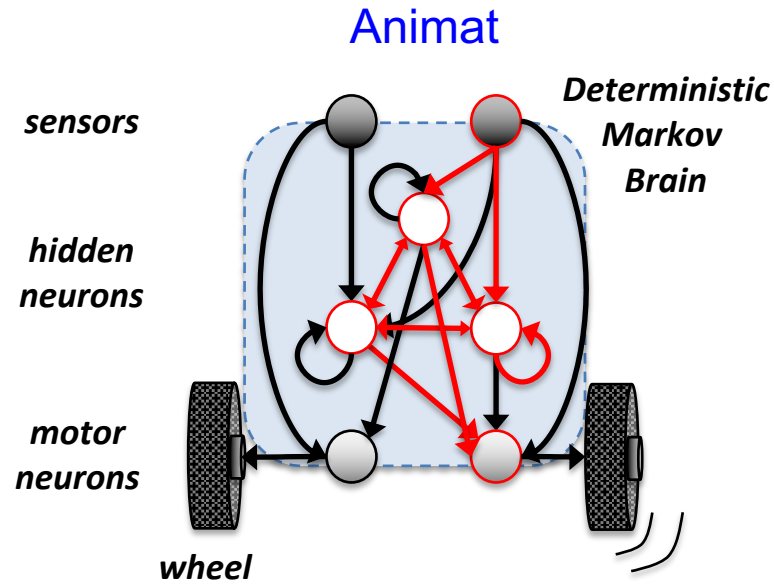




maximum of $\Phi$ in every state in the cycle

Marshall W, Kim H, Walker SI, Tononi G, Albantakis L (2017) How causal analysis can reveal autonomy in models of biological systems. Philos Trans A Math Phys Eng Sci 375:20160358.

# Evolved artificial agents

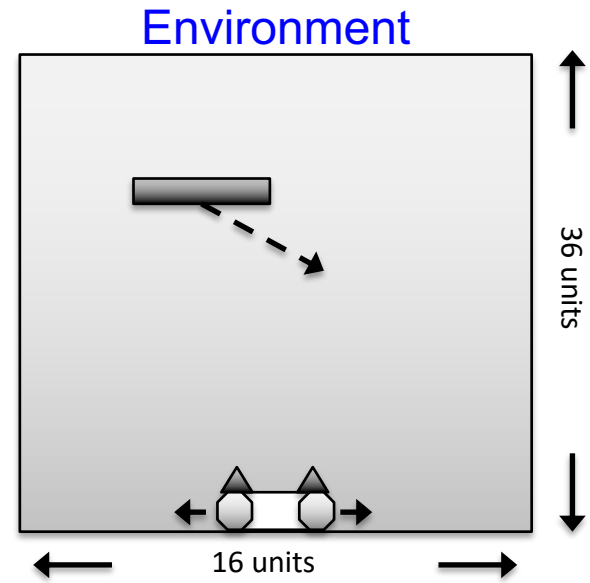


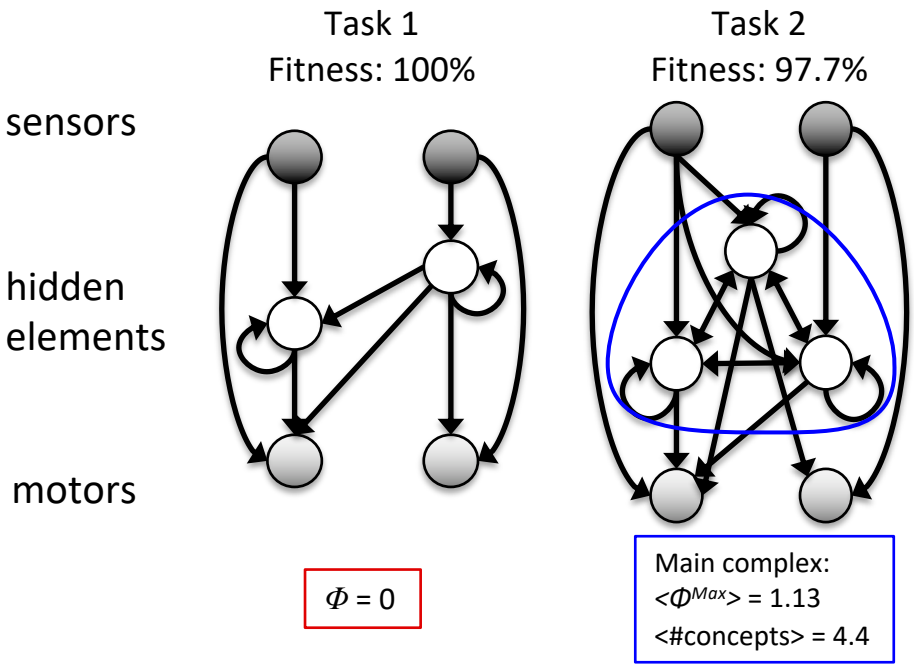

## Selection & Mutation

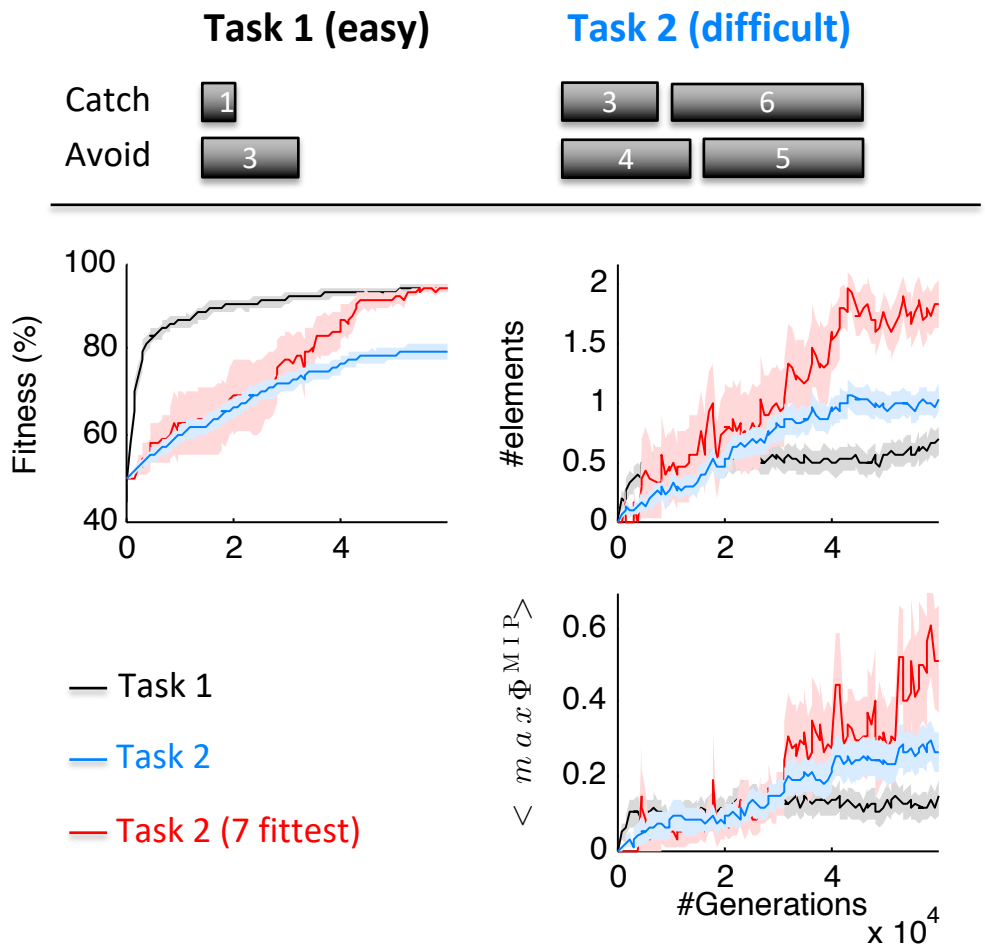
**Fitness**: % of successfully caught and avoided blocks (out of 128 trials)

**Point mutations**, **deletions**, and **duplications** in the genome after each selection



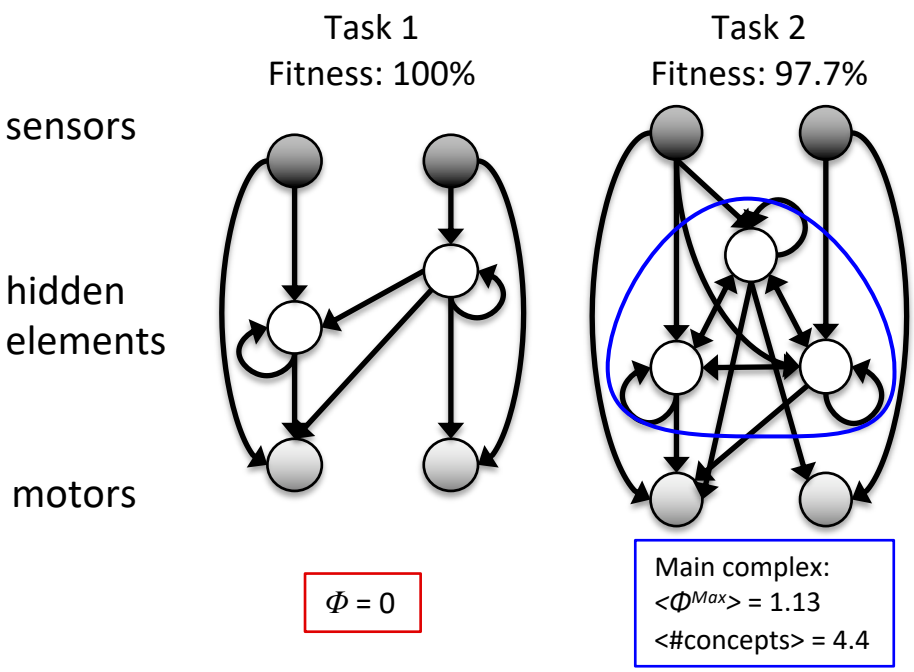
# Larger and more integrated complexes evolved in tasks that require more memory and context-dependent behavior

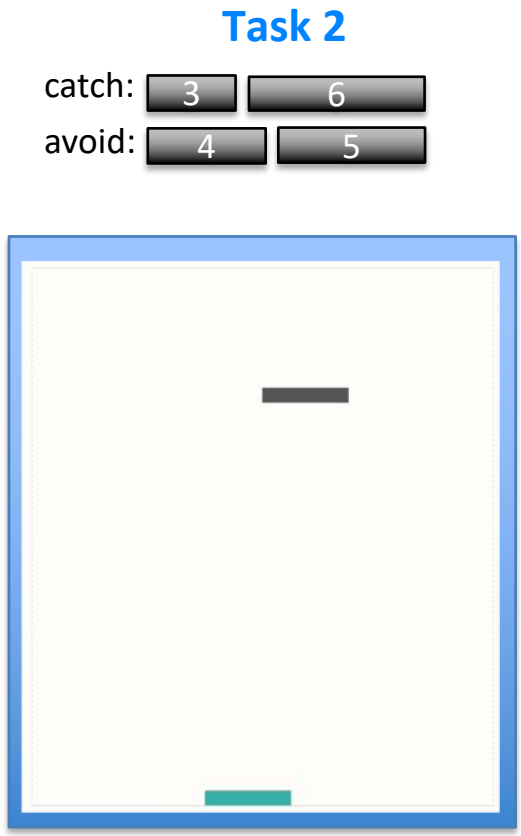



Albantakis L, Hintze A, Koch C, Adami C, Tononi G (2014) Evolution of Integrated Causal Structures in Animats Exposed to Environments of Increasing Complexity. PLoS Comput Biol 10:e1003966.

# Larger and more integrated complexes evolved in tasks that require more memory and context-dependent behavior




Animat is partially autonomous due to memory.

Albantakis L, Hintze A, Koch C, Adami C, Tononi G (2014) Evolution of Integrated Causal Structures in Animats Exposed to Environments of Increasing Complexity. PLoS Comput Biol 10:e1003966.

An autonomous agent is:
1. an open system with stable, self-defined and self-maintained causal borders,
**2. with the capacity to perform actions that are (partially) caused from within**

# 2. Identifying the actual causes of actions

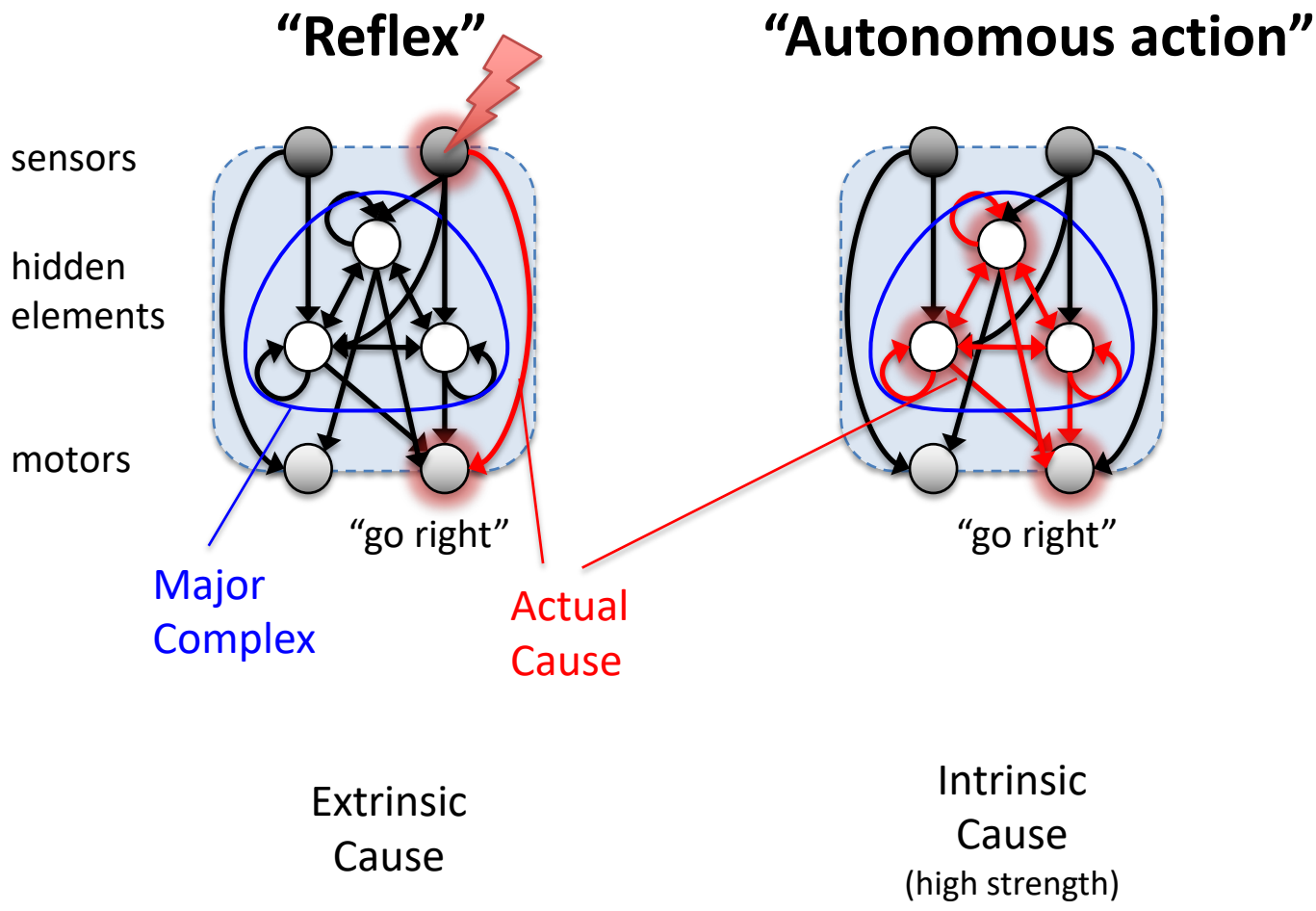
Example 1: Tracing back the causal chain

Albantakis L, Massari F, Tononi G (in prep)

Example 2: Causes of actions across various evolutionary environments

Juel BE, Comolatti R, Tononi G, Albantakis L (2019)

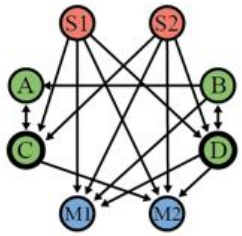
# Extrinsic vs. intrinsic actual causes



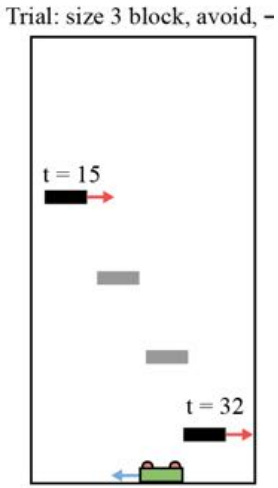

Albantakis L, Marshall W, Hoel E, Tononi G (2019) What caused what? A quantitative account of actual causation using dynamical causal networks. Entropy 21:459.

# Causal analysis: direct actual causes

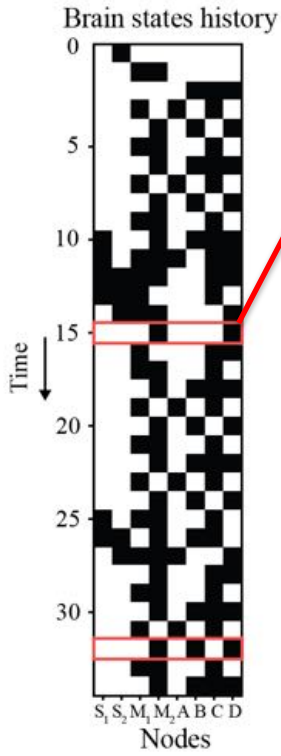
**MB Structure**



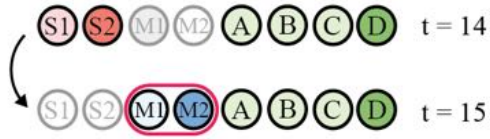

**Task environment**

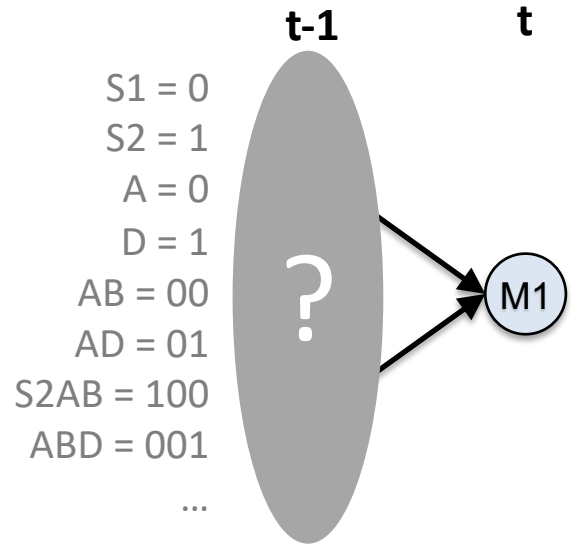



**Time series one trial**




Transition:




Actual cause of $M1_t = 0$?

**Causal strength:**

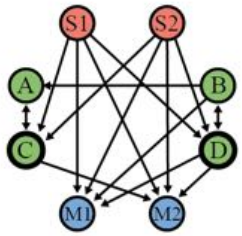
$$\alpha = \log_2\left(\frac{\hat{p}(x_{t-1}|y_t)}{\hat{p}(x_{t-1})}\right)$$

$\hat{p}$: Interventional probabilities
Obtained by setting the elements at t-1 into all of their possible state ("counterfactuals")

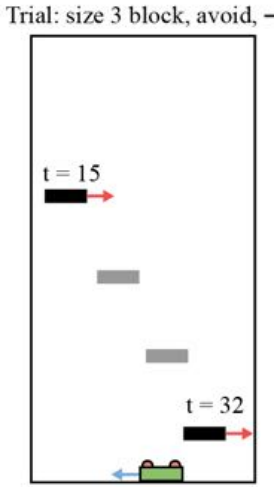
Albantakis L, Marshall W, Hoel E, Tononi G (2019) What caused what? A quantitative account of actual causation using dynamical causal networks. Entropy 21:459.

# Causal analysis: direct actual causes

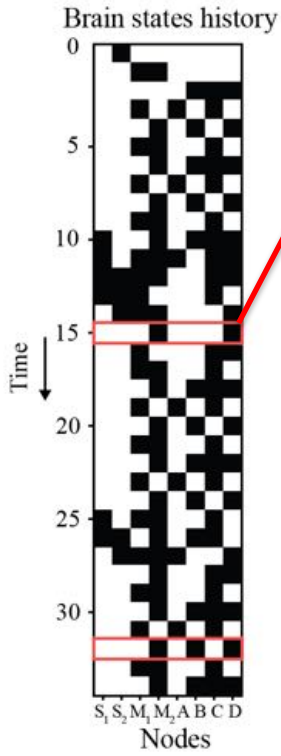
**MB Structure**



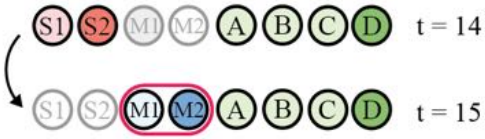

**Task environment**

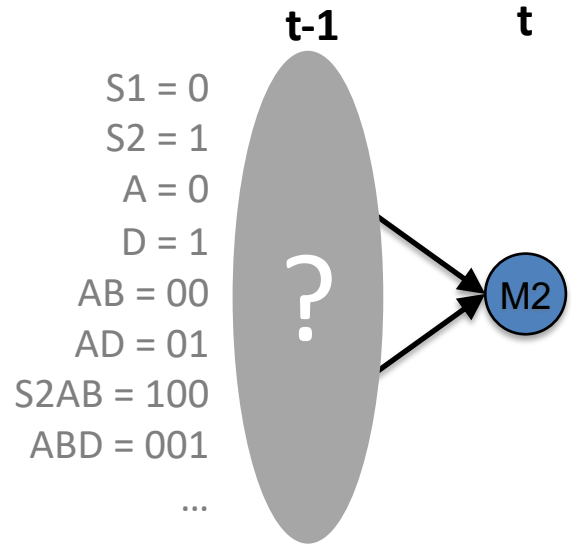



**Time series one trial**




Transition:




Actual cause of $M2_t = 1$?

**Causal strength:**

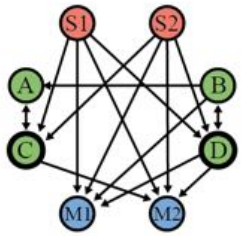
$$\alpha = \log_2\left(\frac{\hat{p}(x_{t-1}|y_t)}{\hat{p}(x_{t-1})}\right)$$

$\hat{p}$: Interventional probabilities
Obtained by setting the elements at t-1 into all of their possible state ("counterfactuals")

Albantakis L, Marshall W, Hoel E, Tononi G (2019) What caused what? A quantitative account of actual causation using dynamical causal networks. Entropy 21:459.

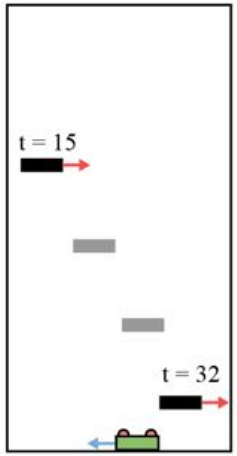
# Causal analysis: direct actual causes

**MB Structure**

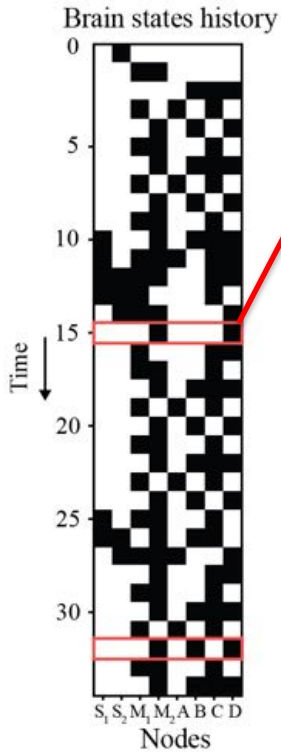



**Task environment**

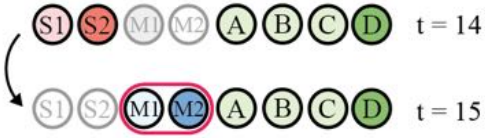
Trial: size 3 block, avoid, –

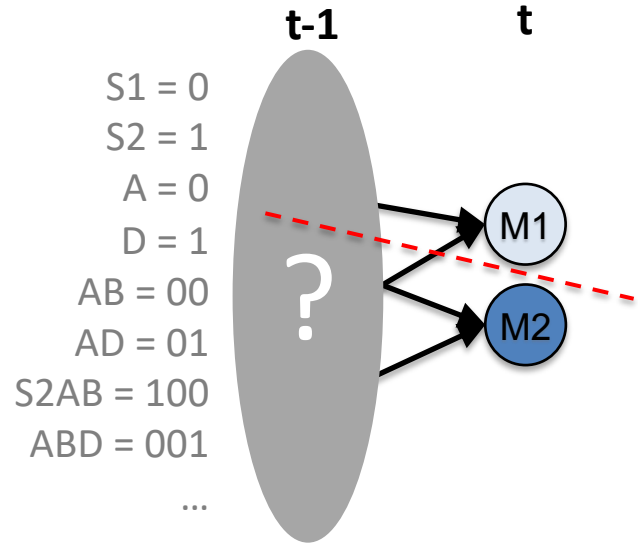



**Time series one trial**




Transition:




Actual cause of $M1M2_t = 01$?

**Causal strength:**

$$\alpha = \min_{\Psi} \left( \log_2 \left( \frac{\hat{p}(x_{t-1}|y_t)}{\Psi(\hat{p}(x_{t-1}|y_t))} \right) \right)$$

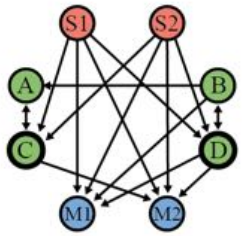
Ψ: partition

$\hat{p}$: Interventional probabilities
Obtained by setting the elements at t-1 into all of their possible state ("counterfactuals")


Albantakis L, Marshall W, Hoel E, Tononi G (2019) What caused what? A quantitative account of actual causation using dynamical causal networks. Entropy 21:459.

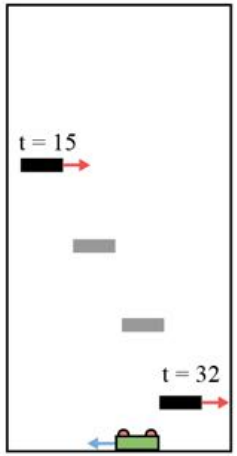
# Same action, same agent, different causes

**MB Structure**

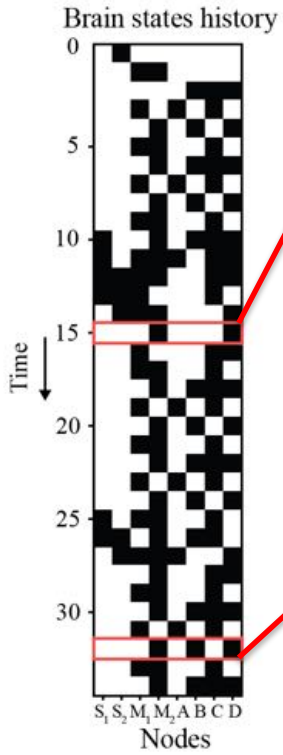



**Task environment**

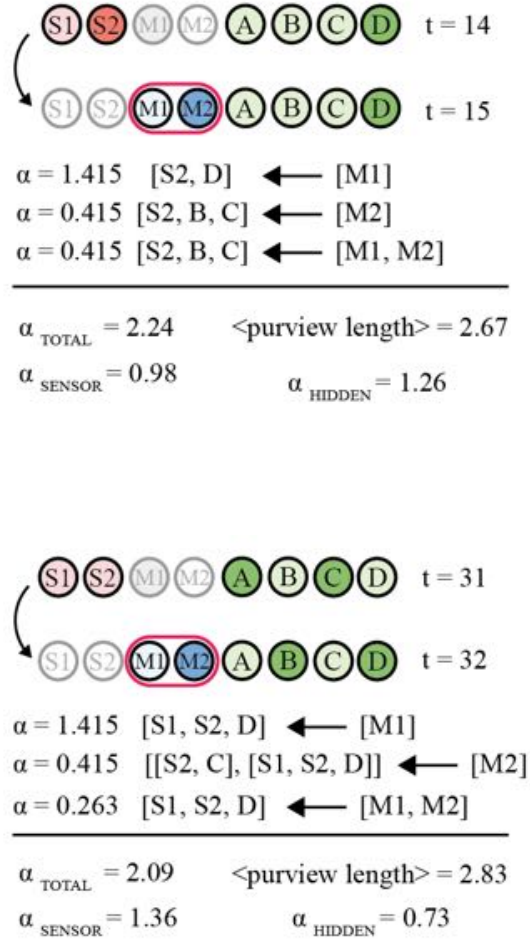
Trial: size 3 block, avoid, –




**Time series one trial**

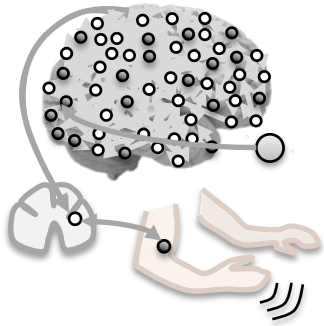



Transition:

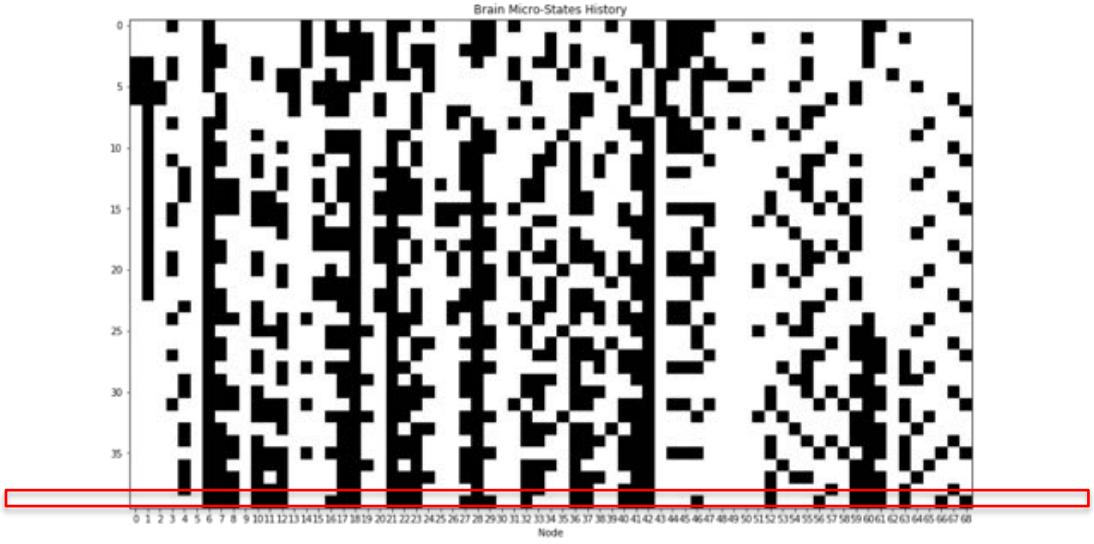


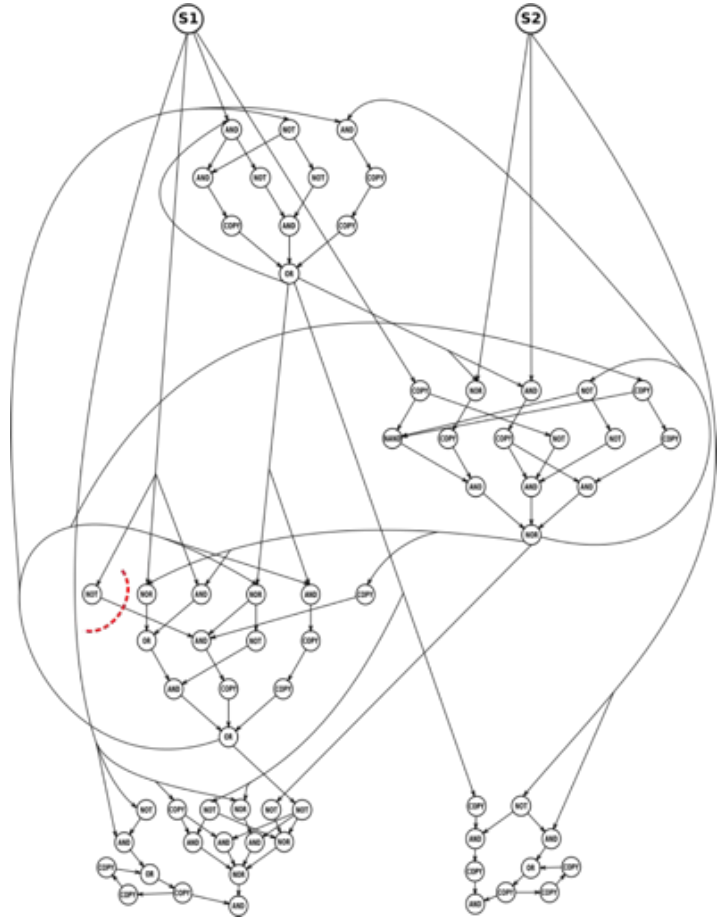

Albantakis L, Marshall W, Hoel E, Tononi G (2019) What caused what? A quantitative account of actual causation using dynamical causal networks. Entropy 21:459.

# Tracing back the causal chain



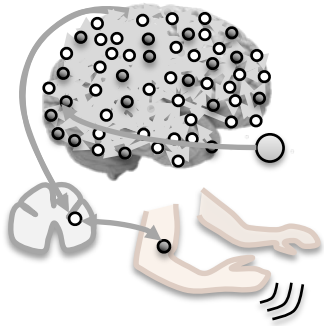
"Neurons were activated and interacted with other neurons which then triggered a motor response."



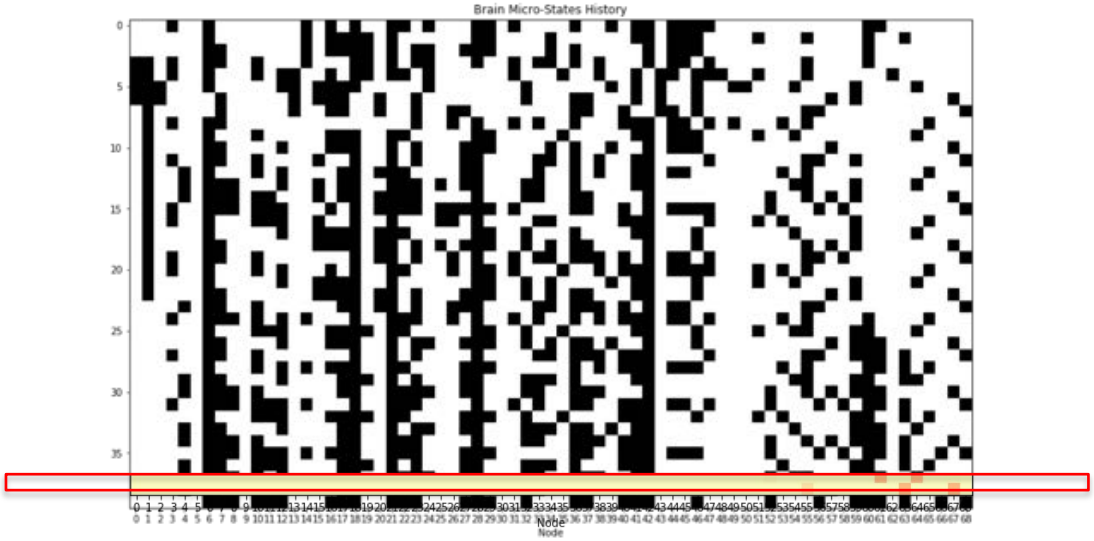



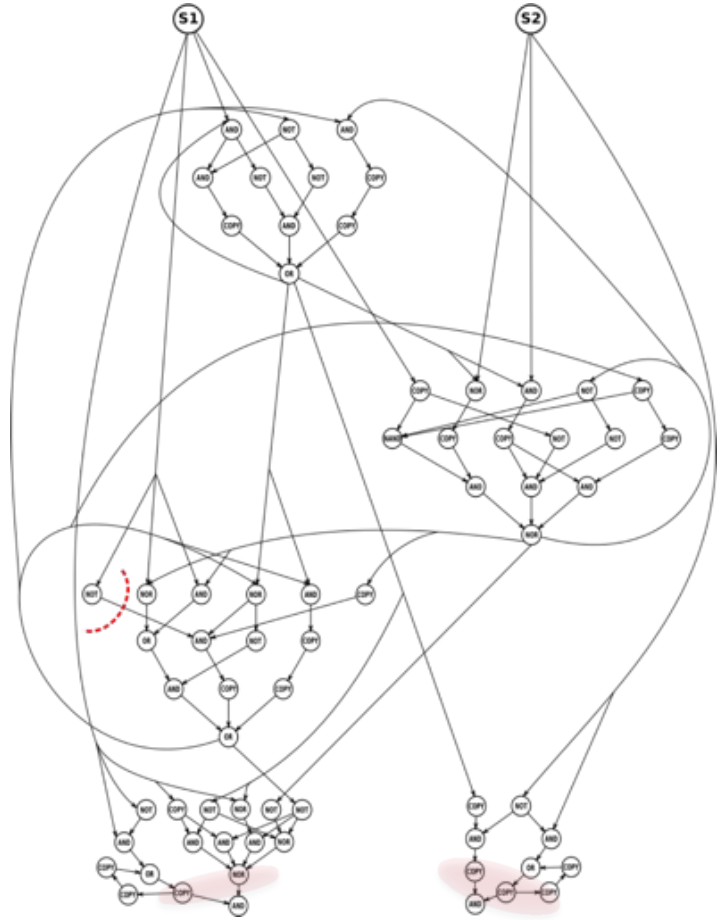

Albantakis L, Marshall W, Hoel E, Tononi G (2019) What caused what? A quantitative account of actual causation using dynamical causal networks. Entropy 21:459.

# Tracing back the causal chain



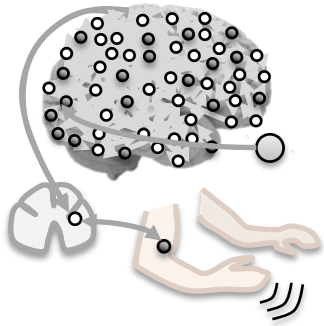
"Neurons were activated and interacted with other neurons which then triggered a motor response."



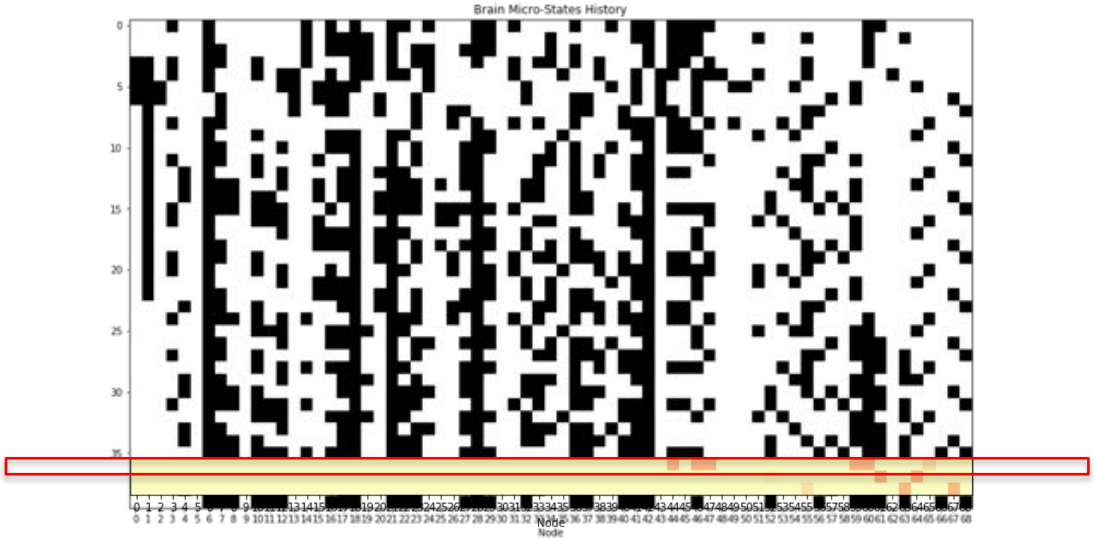



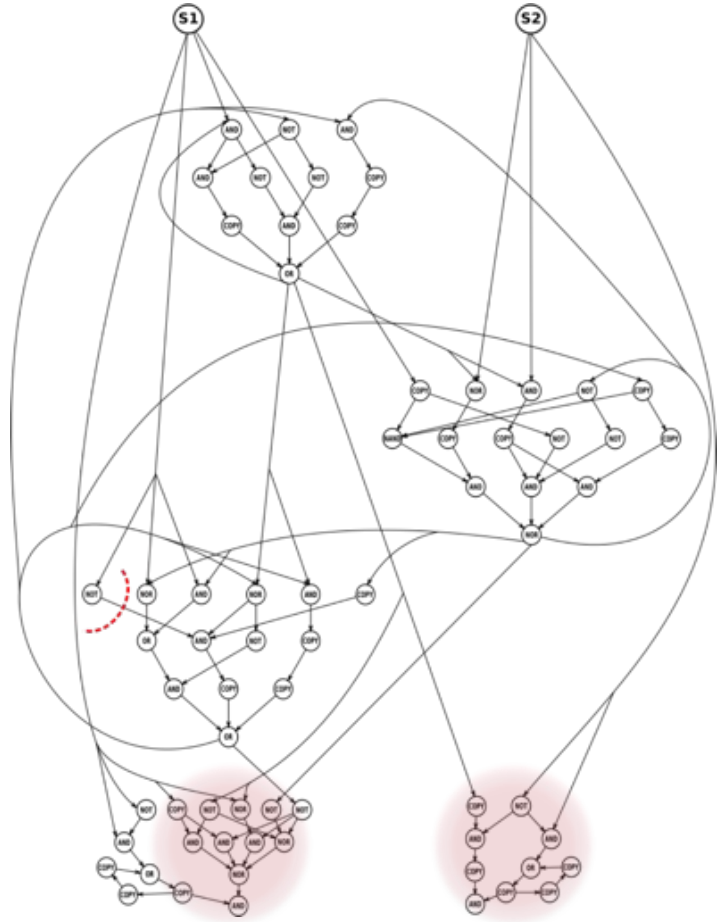

Albantakis L, Marshall W, Hoel E, Tononi G (2019) What caused what? A quantitative account of actual causation using dynamical causal networks. Entropy 21:459.

# Tracing back the causal chain



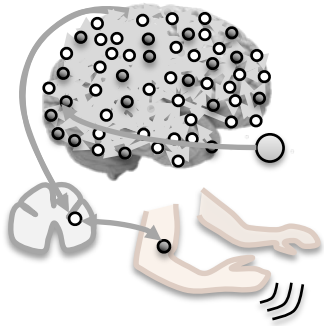
"Neurons were activated and interacted with other neurons which then triggered a motor response."



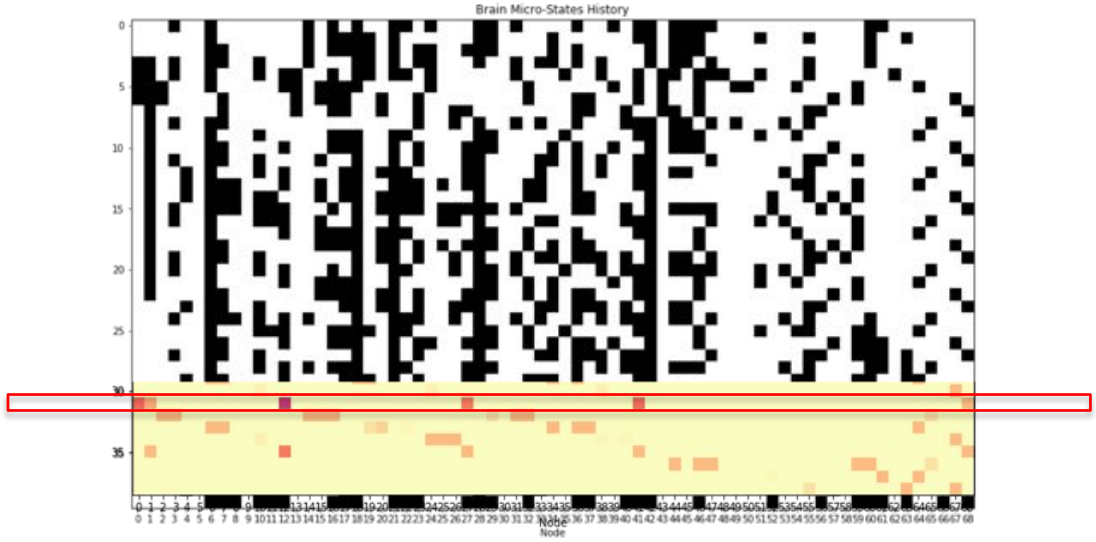



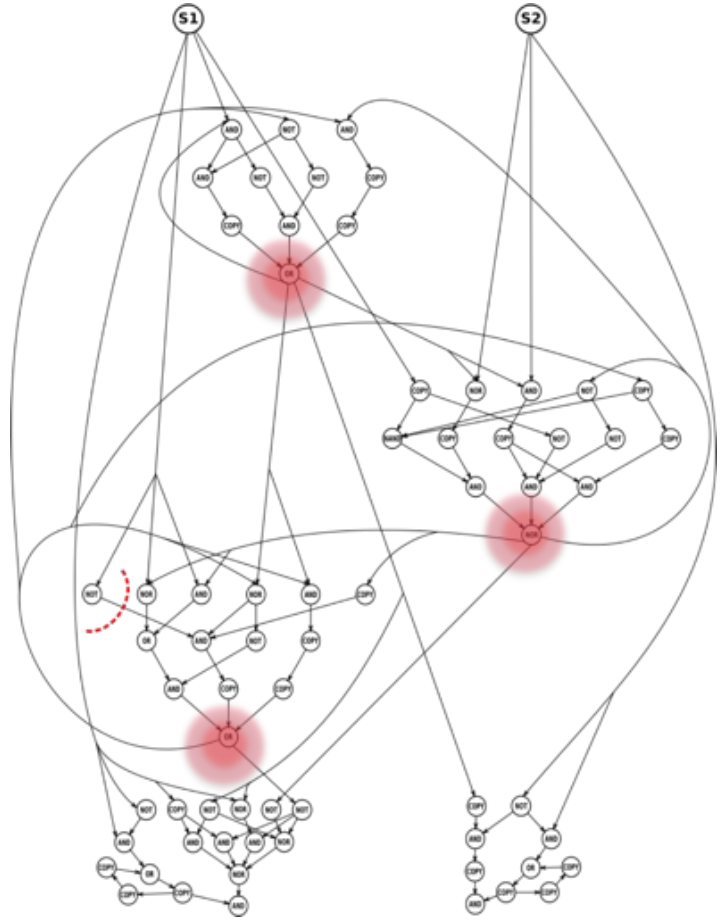

Albantakis L, Marshall W, Hoel E, Tononi G (2019) What caused what? A quantitative account of actual causation using dynamical causal networks. Entropy 21:459.

# Tracing back the causal chain



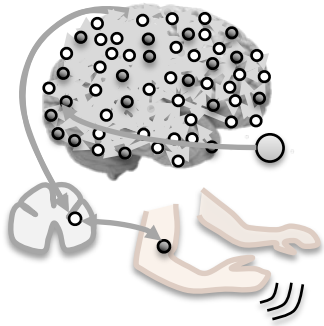
"Neurons were activated and interacted with other neurons which then triggered a motor response."



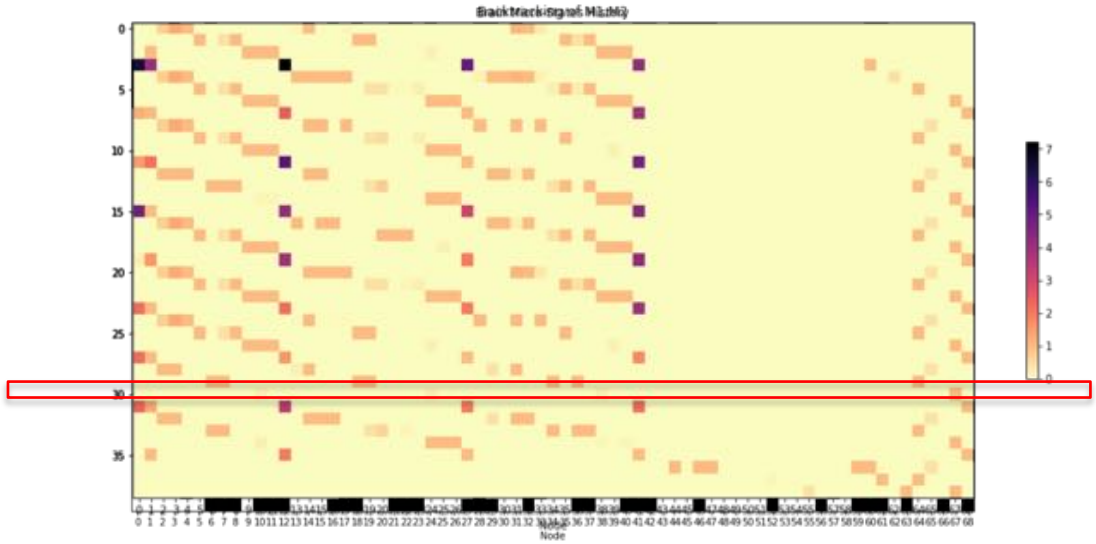



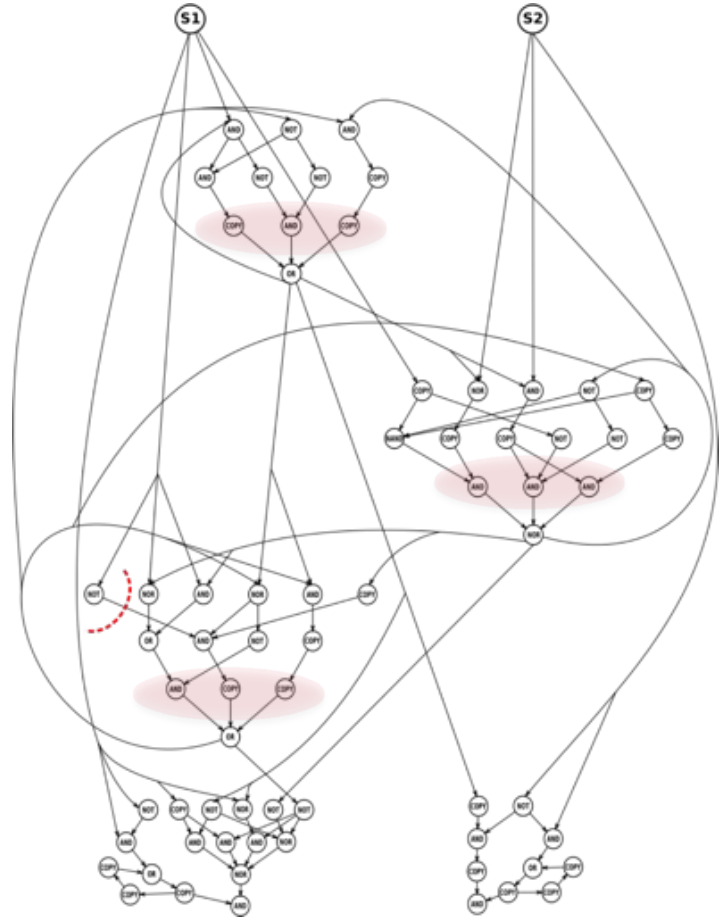

Albantakis L, Marshall W, Hoel E, Tononi G (2019) What caused what? A quantitative account of actual causation using dynamical causal networks. Entropy 21:459.

# Tracing back the causal chain



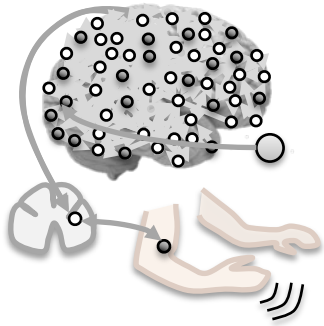
"Neurons were activated and interacted with other neurons which then triggered a motor response."



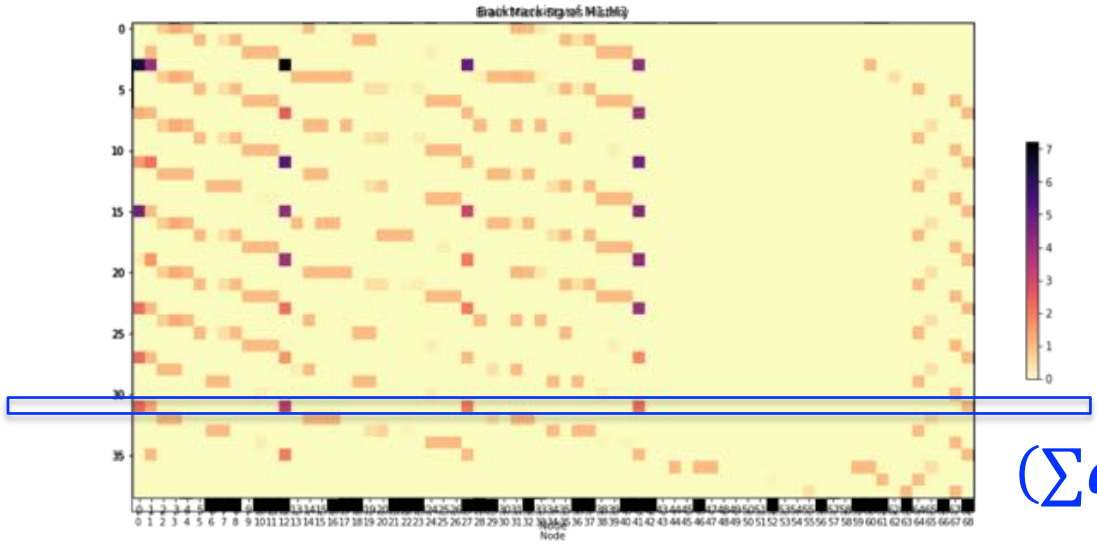



Albantakis L, Marshall W, Hoel E, Tononi G (2019) What caused what? A quantitative account of actual causation using dynamical causal networks. Entropy 21:459.

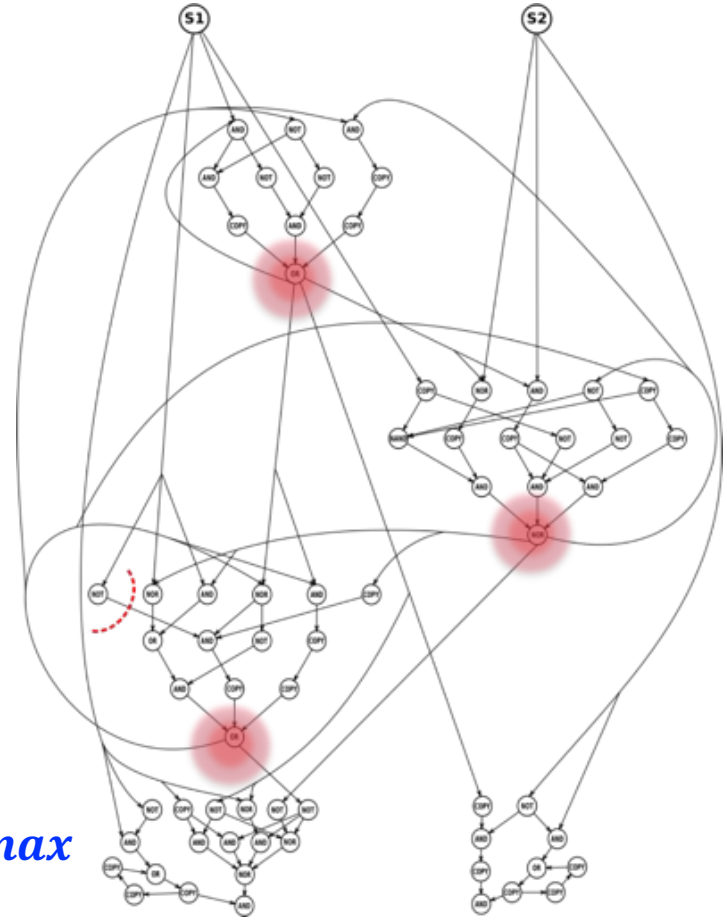
# Tracing back the causal chain



"Neurons were activated and interacted with other neurons which then triggered a motor response."

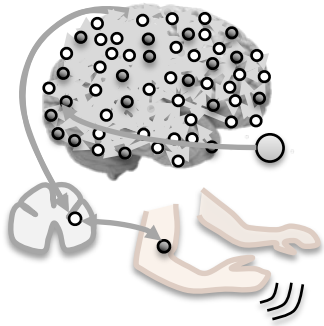


$(\sum \alpha)^{max}$

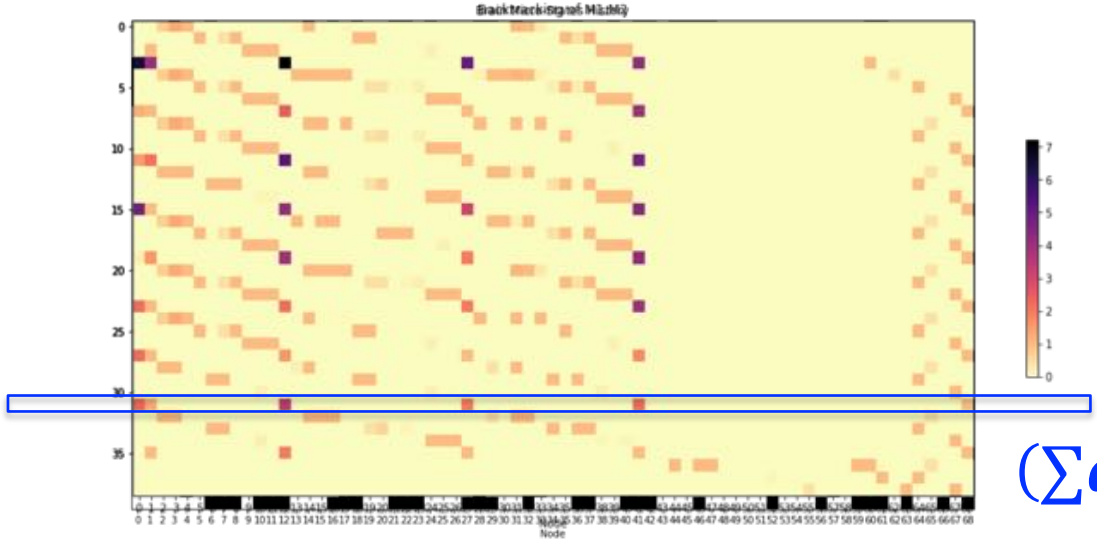



Albantakis L, Marshall W, Hoel E, Tononi G (2019) What caused what? A quantitative account of actual causation using dynamical causal networks. Entropy 21:459.

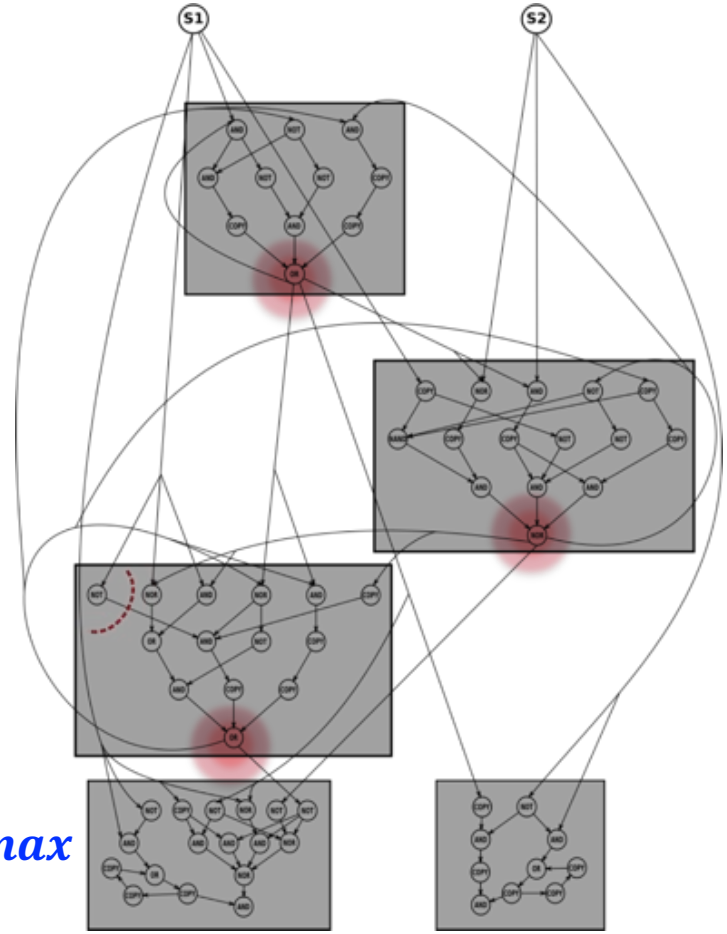
# Tracing back the causal chain



"Neurons were activated and interacted with other neurons which then triggered a motor response."

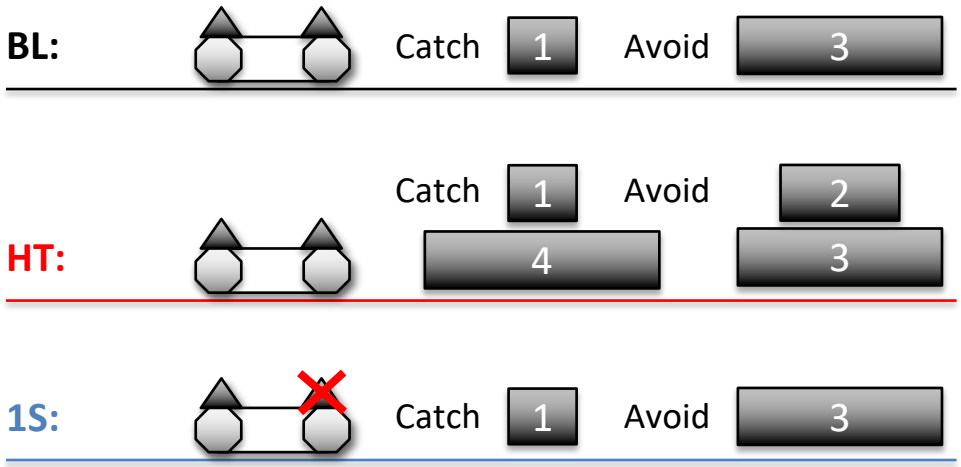


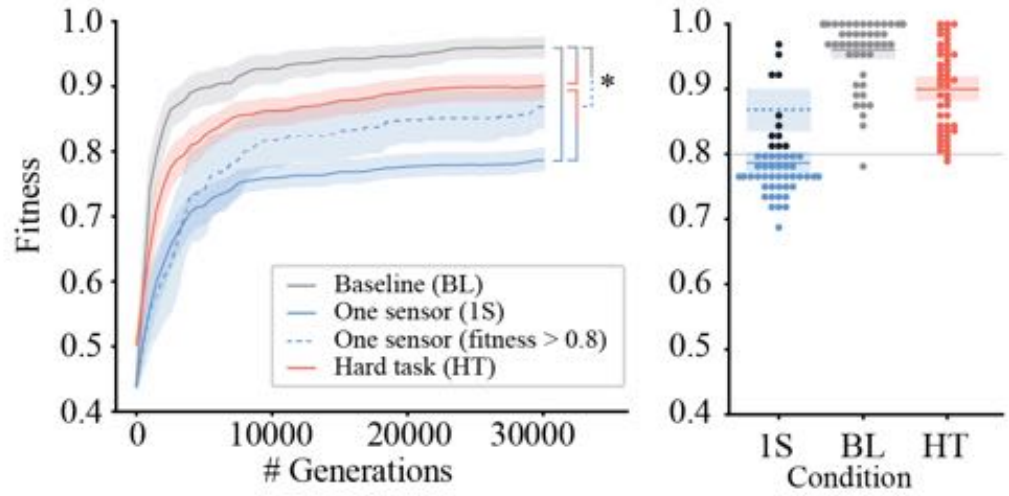
$(\sum \alpha)^{max}$




Albantakis L, Marshall W, Hoel E, Tononi G (2019) What caused what? A quantitative account of actual causation using dynamical causal networks. Entropy 21:459.

# Three different evolutionary environments



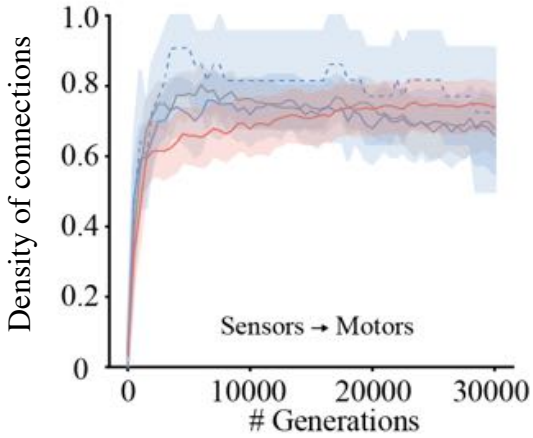



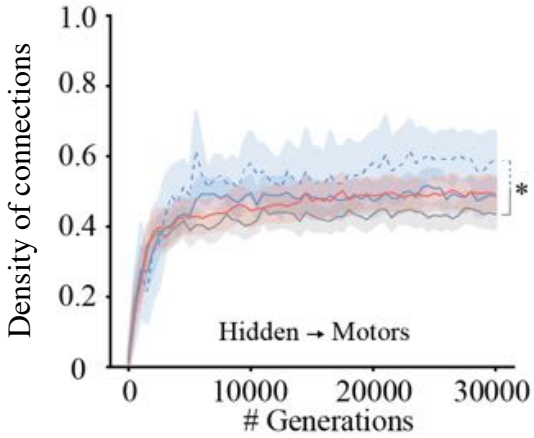

Relative task difficulty: BL < HT < 1S

Juel BE, Comolatti R, Tononi G, Albantakis L (2019) When is an action caused from within? Quantifying the causal chain leading to actions in simulated agents. Arxiv 1904.02995.

# No difference in motor connectivity across conditions

Structural properties:

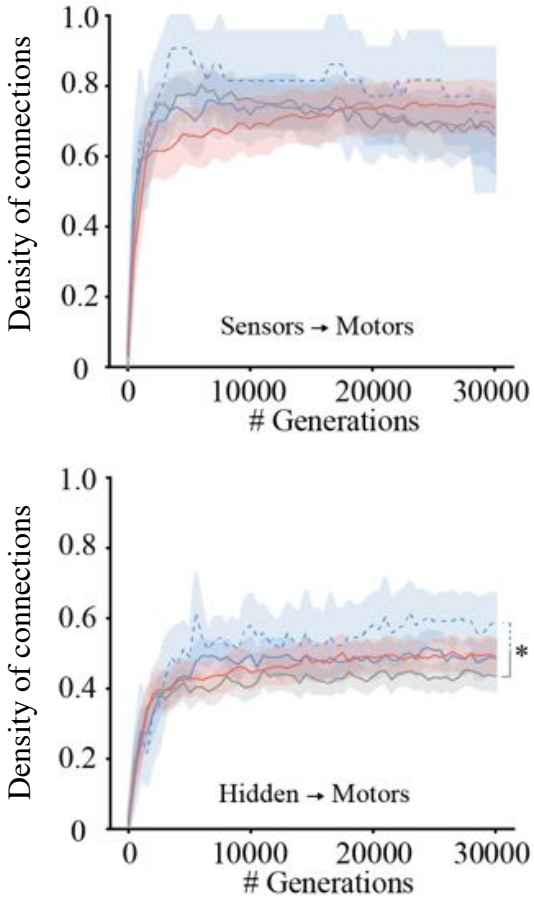


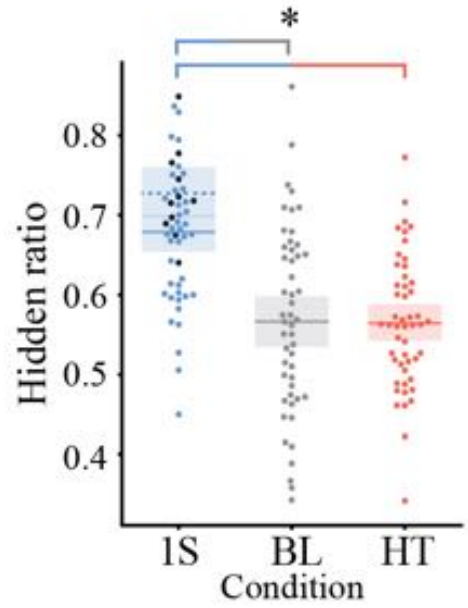

Juel BE, Comolatti R, Tononi G, Albantakis L (2019) When is an action caused from within? Quantifying the causal chain leading to actions in simulated agents. Arxiv 1904.02995.

# Backtracking analysis reveals differences between evolutionary conditions

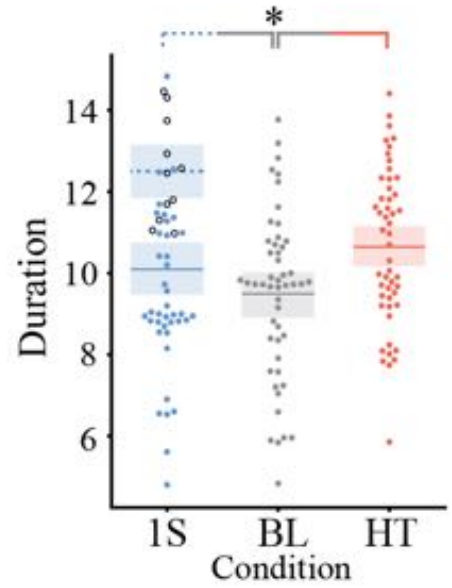
Structural properties:




Backtracking:




1S: More intrinsic




More memory and involvement of hidden states in 1S and HT.

Juel BE, Comolatti R, Tononi G, Albantakis L (2019) When is an action caused from within? Quantifying the causal chain leading to actions in simulated agents. Arxiv 1904.02995.

An autonomous agent is:
1. an open system with stable, self-defined and self-maintained causal borders,
2. with the capacity to perform actions that are (partially) caused from within

# 1. Identifying self-defined causal borders

Example 1: Fission yeast cell cycle model

Marshall W, Kim H, Walker SI, Tononi G, Albantakis L (2017)

Example 2: Evolved artificial agents (animats)

Albantakis L, Hintze A, Koch C, Adami C, Tononi G (2014)

# 2. Identifying the actual causes of actions

Example 1: Tracing back the causal chain

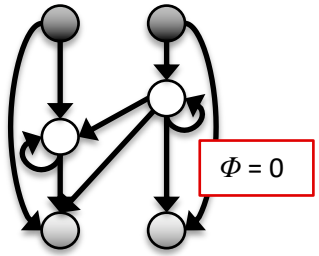
Albantakis L, Massari F, Tononi G (in prep)

Example 2: Causes of actions across various evolutionary environments

Juel BE, Comolatti R, Tononi G, Albantakis L (2019)

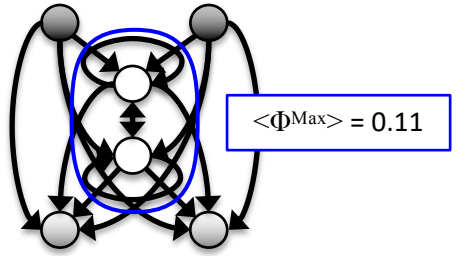
# Two dissociations

Feed-forward
+ self-loop network



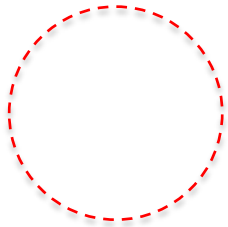

Generation
#59904

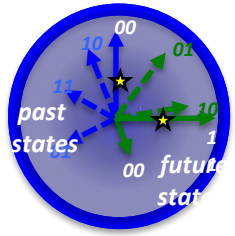
Feed-back network




Generation
#59904

100% Fitness





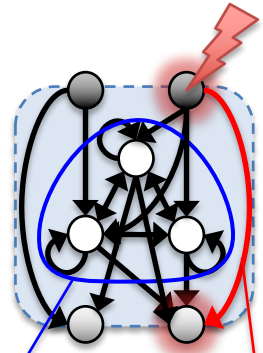

"Reflex"

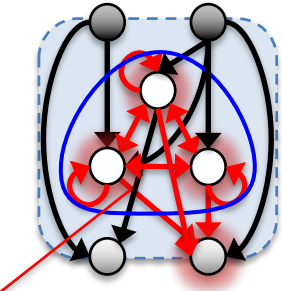
"Autonomous action"

sensors

hidden
elements

motors





"go right"

"go right"

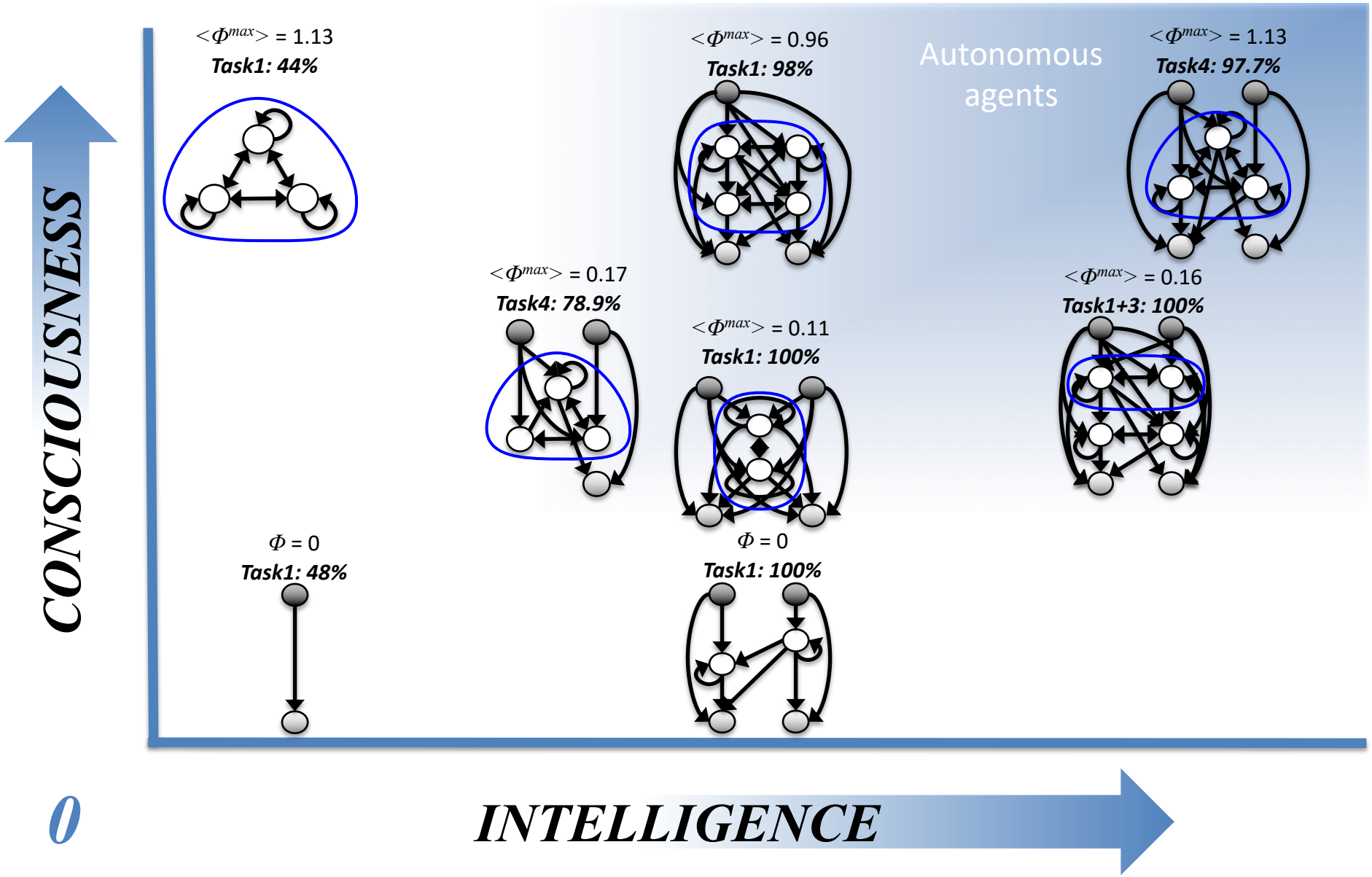
Major
Complex

Actual
Cause

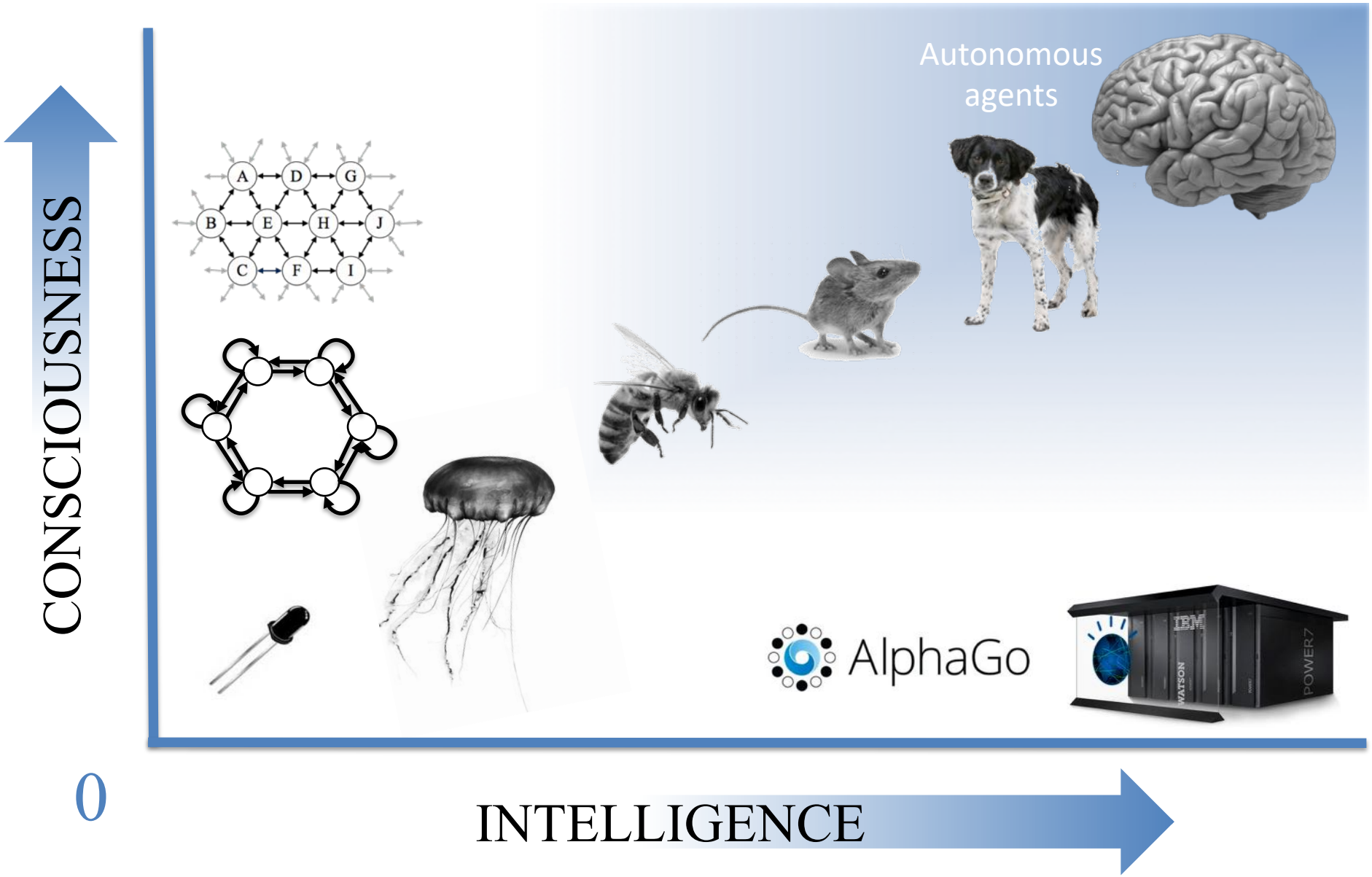
Extrinsic
Cause

Intrinsic
Cause
(high
strength)

# Intelligence vs. Consciousness



# Intelligence vs. Consciousness



# Thank you!

http://integratedinformationtheory.org



Giulio Tononi
William Marshall
Bjørn Juel
Renzo Comolatti
Will Mayner