Intrinsic Causation and Consciousness

Larissa Albantakis[1*], Erik P. Hoel[1], Masafumi Oizumi[1,2], Christof Koch[3], and Giulio Tononi[1]

[1] Department of Psychiatry, University of Wisconsin, Madison, Wisconsin, USA
[2] RIKEN Brain Science Institute, Wako-shi, Saitama, Japan
[3] Allen Institute for Brain Science, Seattle, Washington, USA
*Presenting author: albantakis@wisc.edu

Abstract

Building upon integrated information theory (IIT) (e.g. Tononi, 2012), we investigate parallels between phenomenological axioms of consciousness (existence, compositionality, information, integration, and exclusion) and the intrinsic causal structure of physical systems.

Central to our approach is the claim that causation has an *informational* aspect: a mechanism can only have a causal role within the system, if its present state constrains the potential past and future of the system, referenced to all possible counterfactuals (system states). Sets of mechanisms can have a *compositional* causal role, provided they are *integrated*, meaning they are irreducible to the causal roles of their parts. By assessing the amount of integrated information specified by sets of mechanisms, one can assess both the quality, or 'causal role', and the quantity, or 'causal power' they exert within a system. Finally, sets of mechanisms that generate local maxima of irreducible causal information form causal *complexes*. These exclude other, overlapping causal entities, and thereby avoid overdetermination.

These causal principles are illustrated in simple networks of neuron-like linear threshold units. Our approach has consequences not only for the fundamental relationship between information and causation, but also for our understanding of emergence, adaptation, and meaning. Finally, this approach identifies consciousness with a local maximum of causation (a causal complex).

Introduction

The causal interactions in a system are typically evaluated from the extrinsic perspective of an observer. Here, we take the intrinsic perspective of the system itself and ask:

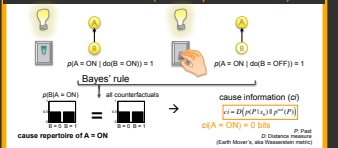
- How can we quantify intrinsic causation in a system of mechanisms (such as the brain)?
- What are the causal roles (***concepts***) of individual and compositional mechanisms?
- When is a system a causal entity (***complex***) from its intrinsic perspective?

By identifying the requisites to assess intrinsic causation (causation within the system from the point of view of the system), we arrive at the same principles postulated by the integrated information theory (IIT) of consciousness, which are derived from phenomenological axioms about conscious experience: information, integration, and exclusion.

Mechanisms: Information

Measuring differences that make a difference to a system, from the intrinsic perspective of the system

Counterfactuals (causal perturbations)

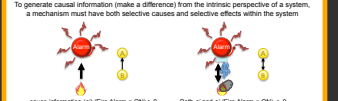


Selectivity (information)

Causes and effects (within the system)

To generate causal information (make a difference) from the intrinsic perspective of a system, a mechanism must have both selective causes and selective effects within the system

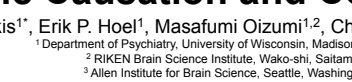
Cause-effect information



Mechanisms: Integration

From the intrinsic perspective of a system, only *irreducible* information matters

(when the whole generates more information than the parts)



Irreducibility is assessed by causally partitioning elements, i.e. by 'noising' connections

Irreducible causes: **MIP: minimum information partition** Irreducible effects:

$\varphi^{MIP}(P|x_0) = D\left(\left(P|x_0\right), \prod\left(P|x_0/MIP\right)\right)$ $\quad \varphi^{MIP}(F|x_0) = D\left(\left(F|x_0\right), \prod\left(F|x_0/MIP\right)\right)$

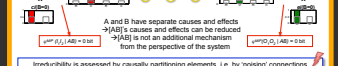
Integrated causes and effects

Only cause or effect repertoires are irreducible

$\varphi^{MIP} = \min\left(\varphi^{MIP}\left(P|x_0\right), \varphi^{MIP}\left(F|x_0\right)\right)$

A candidate concept

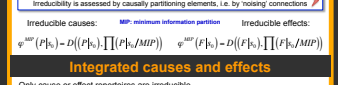
Both cause and effect repertoires are irreducible: A candidate **concept**



Mechanisms: Exclusion

To avoid **overcausation**, a mechanism can specify only one concept: the cause-effect repertoire that is maximally irreducible (φ_{max}^{MIP})

$\varphi_{max}^{MIP} = \max\left(\varphi_{max}^{MIP}\left(P|x_0\right), \varphi_{max}^{MIP}\left(F|x_0\right)\right)$

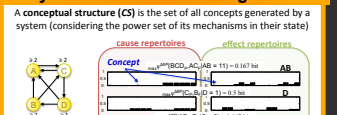


Concept: *a maximally* irreducible cause-effect repertoire (core IF-THEN function) from the intrinsic perspective of a mechanism; all other functions are excluded (treated as extrinsic "noise")

System: Information/Integration

A **conceptual structure** (***CS***) is the set of all concepts generated by a system (considering the power set of its mechanisms in their state)

Integrated conceptual structure



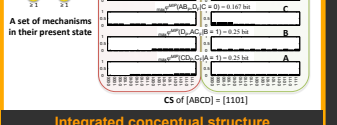
Irreducibility is assessed by unidirectional partitions (strong integration): each element must be able to affect / be affected by every other element

Integrated information

$\Phi^{MIP} \rightarrow \left(C|x_0\right) = D\left(\left(C|x_0\right), \left(C|x_0/MIP \rightarrow\right)\right)$

MIP: minimum information partition

A candidate complex

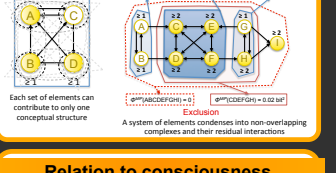


System: Exclusion

Complex: a subset of elements within a system generating a maximally integrated conceptual structure ($_{max}\Phi^{MIP}$)

Relation to consciousness

Consciousness: a local maximum of intrinsic causation ($_{max}\Phi^{MIP}$) over all sets of elements and space/time (Hoel et al., in prep.)



Conclusion

Causation should be assessed using:

1) perturbations 2) all counterfactuals 3) with equal probability

From the intrinsic perspective of a system:

- causation and information are one and the same (differences that make a difference)
- Mechanism and system boundaries have to arise out of the system's causal structure ($_{max}\Phi^{MIP}$ $_{max}\Phi^{MIP}$)

Overcausation can be avoided by causally excluding everything but the most informative causes and effects.

Axioms

Identifying the essential properties of consciousness

0th **Existence:** Consciousness exists

1st **Composition:** Consciousness is structured: every experience has many aspects

2nd **Information:** Consciousness is differentiated: every experience is what it is by differing in its particular way from many other experiences

3rd **Integration:** Consciousness is unified: every experience cannot be reduced to non-interdependent components ('strong' integration)

4th **Exclusion**: Consciousness is exclusive: every experience has definite borders and grain size (over space and time)

Postulates

Requirements for the physical substrate of consciousness

0th **Existence:** Mechanisms in a state exists

1st **Composition:** Elementary mechanisms can be combined into higher-order ones

2nd **Information:** Only mechanisms that specify 'differences that make a difference' within a system count

3rd **Integration:** Only information irreducible to non-interdependent components counts

4th **Exclusion**: Only maxima of integrated information count (over elements, space, time)

Causal information vs. Shannon information

Information (CEI):	**differences that make a difference (causes/effects)** from the **intrinsic** perspective of a **system**
Integration (Φ^{MIP}):	necessary: only **irreducible** interactions matter
Exclusion ($_{max}\Phi^{MIP}$):	**maximally irreducible** interactions supersede others

- Assessed by **perturbation**: by trying all possible counterfactuals with equal likelihood
- Defined for mechanisms in a **single state:** how past and future states are constrained
- Defined for mechanisms (concepts -> distributions) **and** systems of mechanisms (complexes -> structures)
- Both **irreducibility** (quantity) and **shape** (quality) of distributions / structures matter (EMD)

Information (MI):	**signal transmission** across a (noisy) channel (I/O correlations) from the **extrinsic** perspective of an **observer**
Integration:	not necessary: only **inputs and outputs** matter
Exclusion:	no: arbitrary encodings

- Assessed by **observation**: by recording correlations
- Defined for ensemble **averages**: capacity of a channel
- Defined for channels
- Only selectivity (**reduction of uncertainty**) of distributions matters (KL-Distance)

Acknowledgments:
This work has been supported by the DARPA grant HR 0011-10-C-0052 on "Physical Intelligence".

References

Albantakis L, Hoel EP, Oizumi M, Koch C, Tononi G (in prep.) Causation from a system's perspective.
Hoel EP, Albantakis L, Tononi G (in prep.) When Macro beats Micro: Quantifying causal emergence.
Tononi G (2012) Integrated Information Theory of Consciousness: An Updated Account. Arch Ital Biol 150: 56–90.
Oizumi M, Albantakis L, Tononi G (in prep.) Integrated Information Theory of Consciousness – an update.