Evaluating Artificial Consciousness through Integrated Information Theory

Dr. William Marshall (wmarshall@brocku.ca)
Department of Mathematics and Statistics, Brock University
Institute for Noetic Science – How to Conceive of a Conscious AI

Acknowledgements

- Giulio Tononi

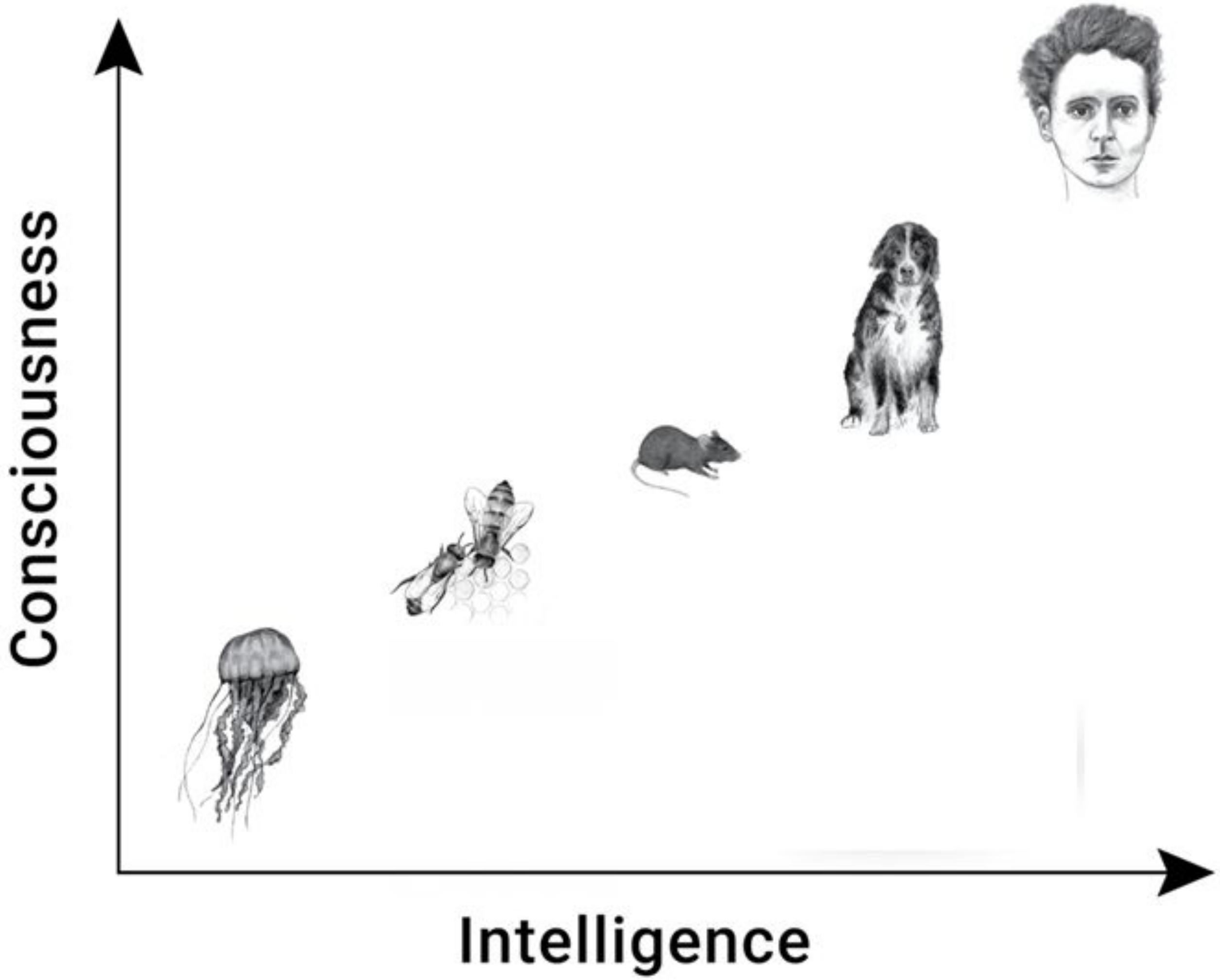
- Larissa Albantakis

- Graham Findlay

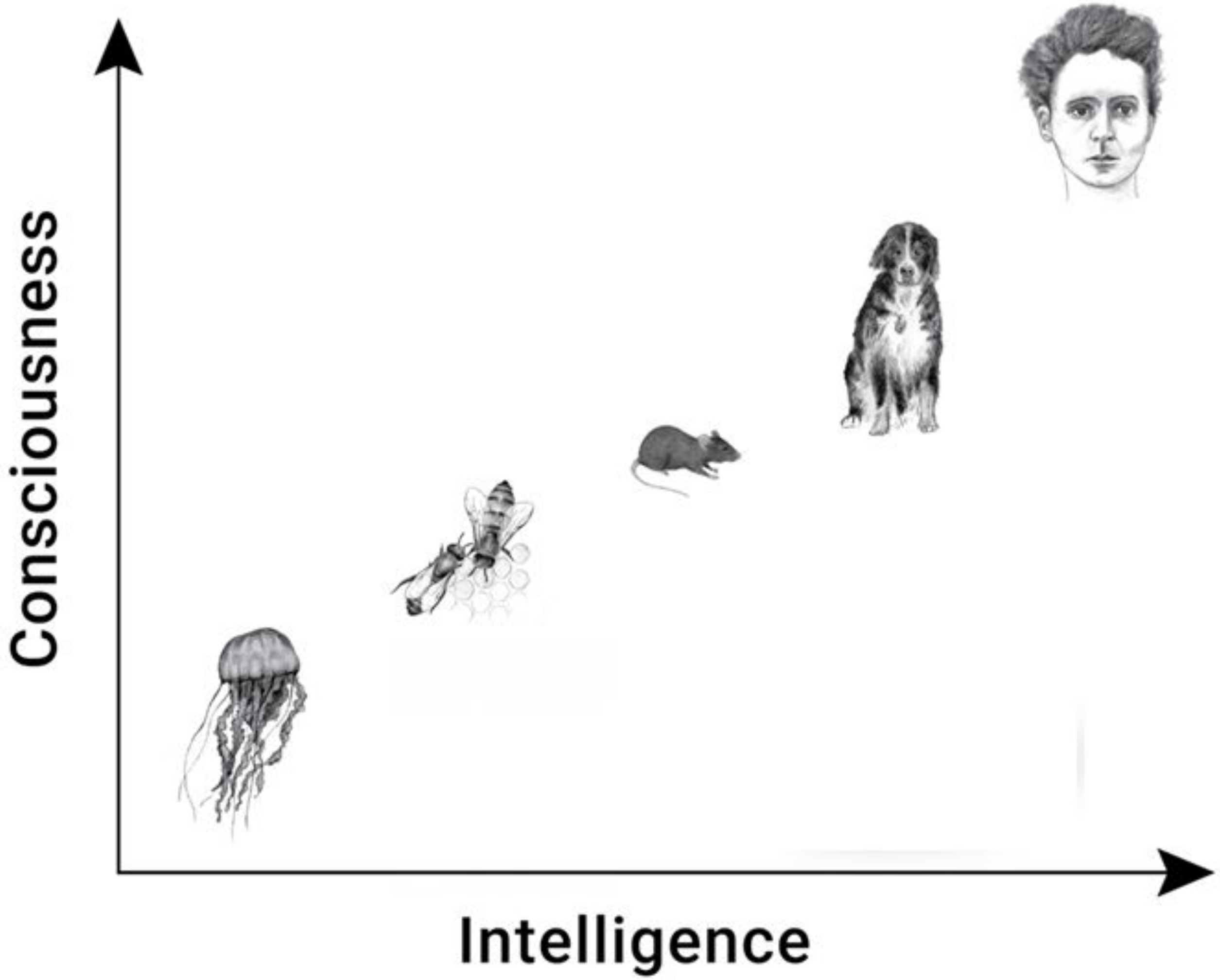
www.iit.wiki

Why study artificial consciousness?

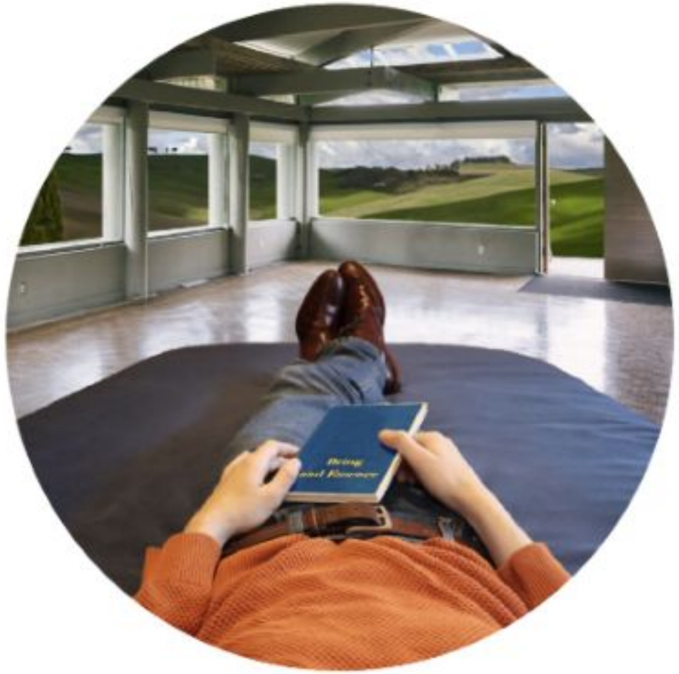
Intelligence and Consciousness?



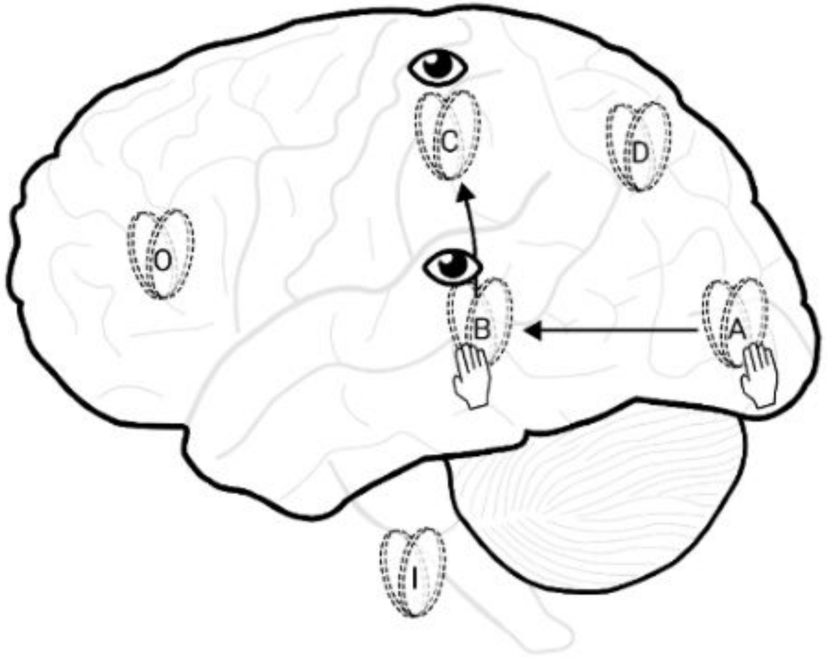
Intelligence and Consciousness ?



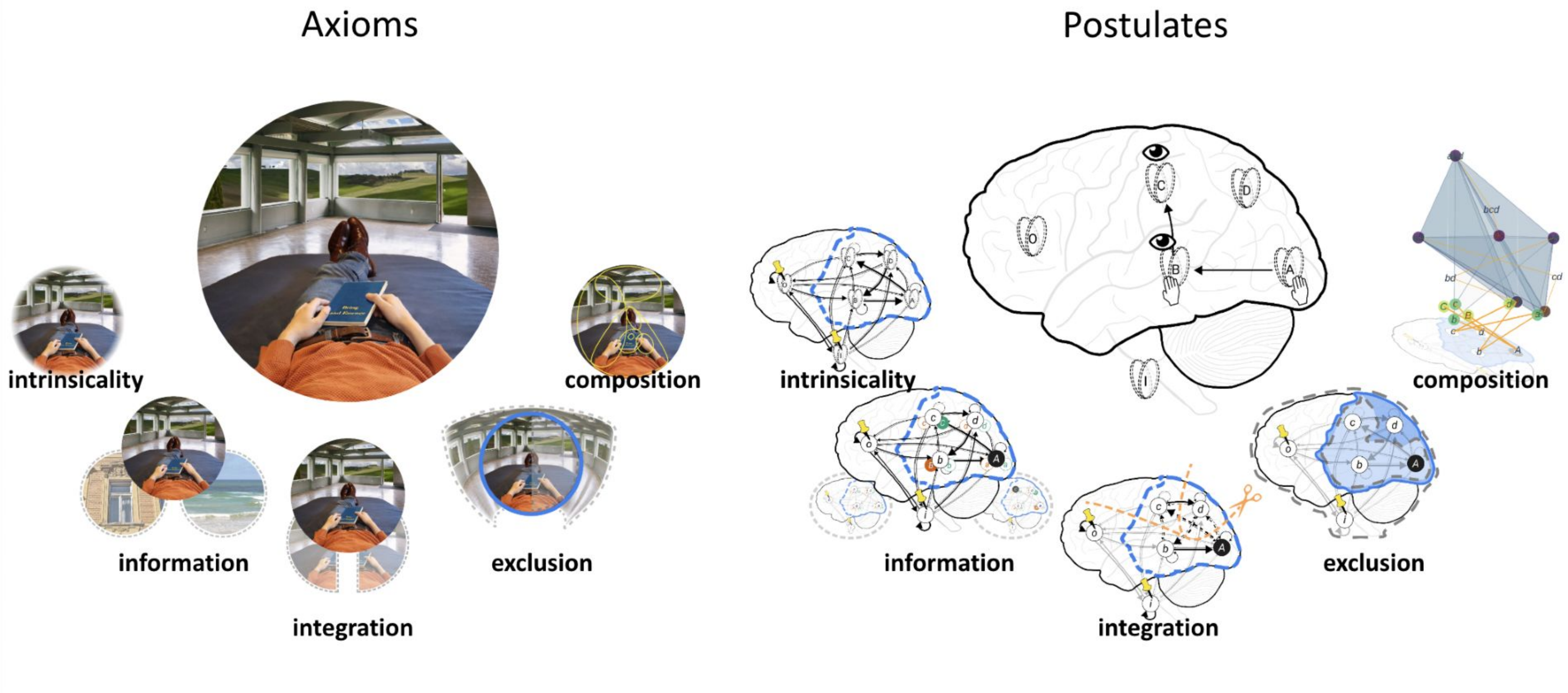
Integrated Information Theory



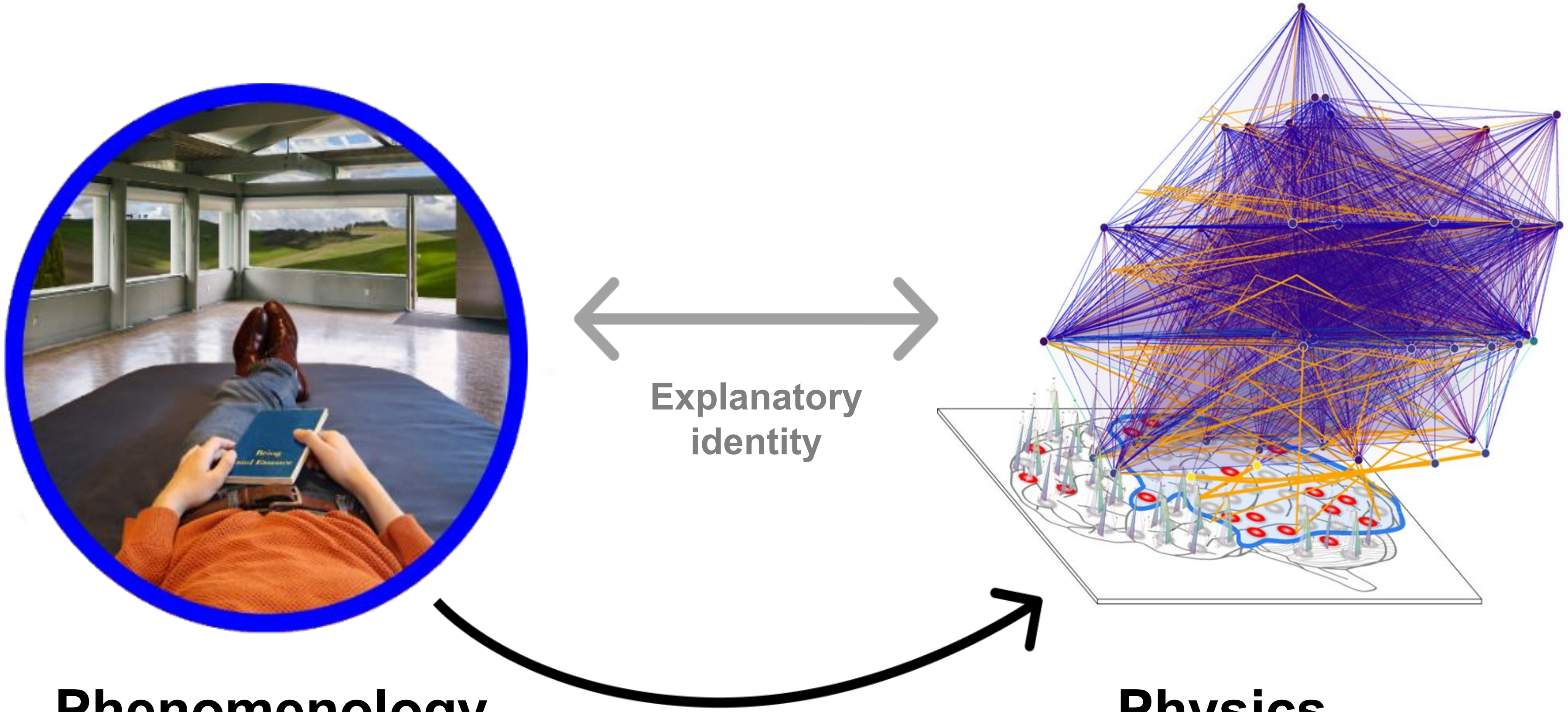
Integrated Information Theory



Integrated Information Theory



Integrated Information Theory

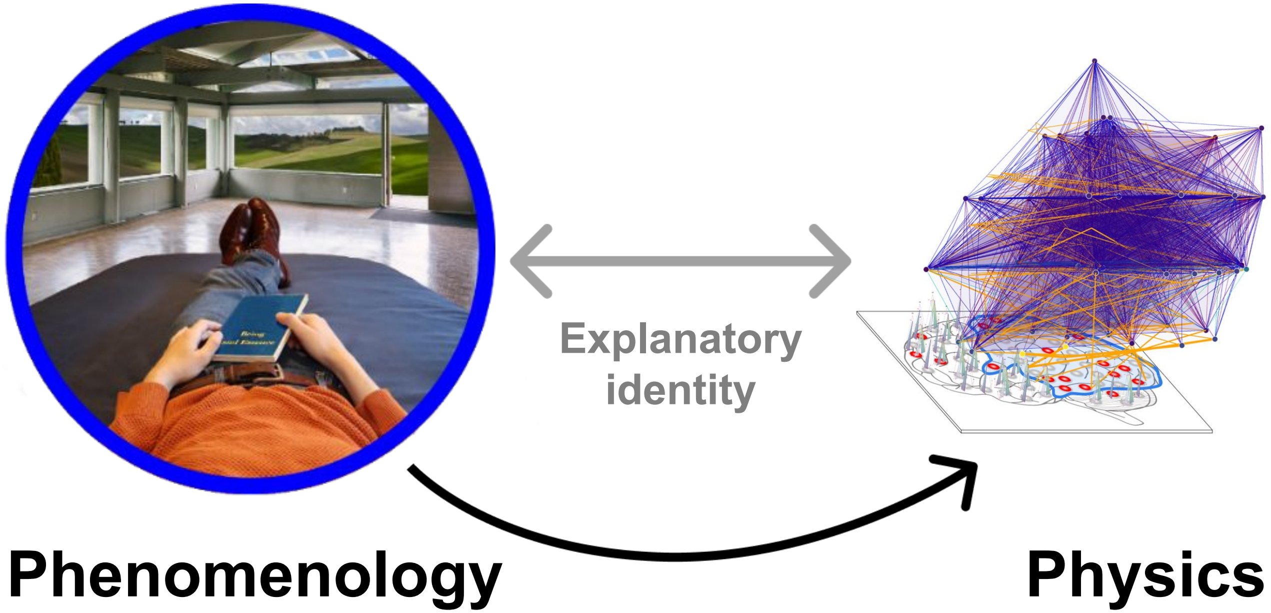


Explanatory identity

Phenomenology

Physics

A Scientific Approach



Explanatory identity

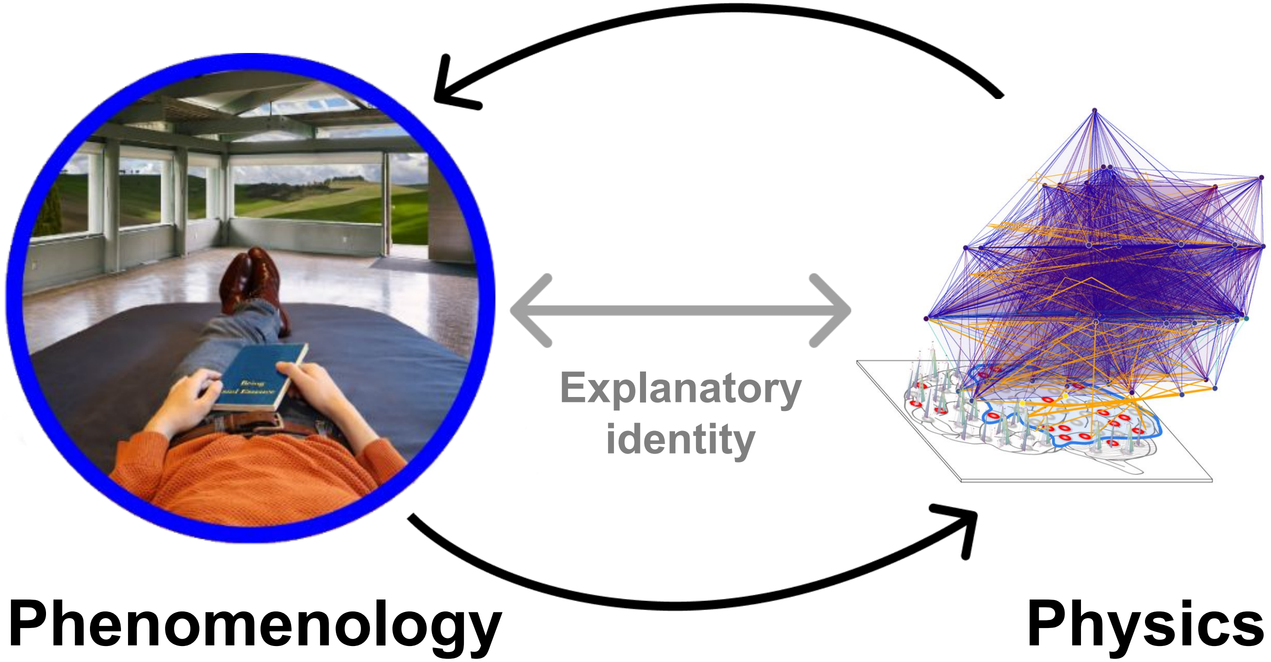
Phenomenology

Physics

How could the properties of my experience be accounted for by the properties of its physical substrate?

A Scientific Approach

Is this thing conscious? In what way?



Explanatory identity

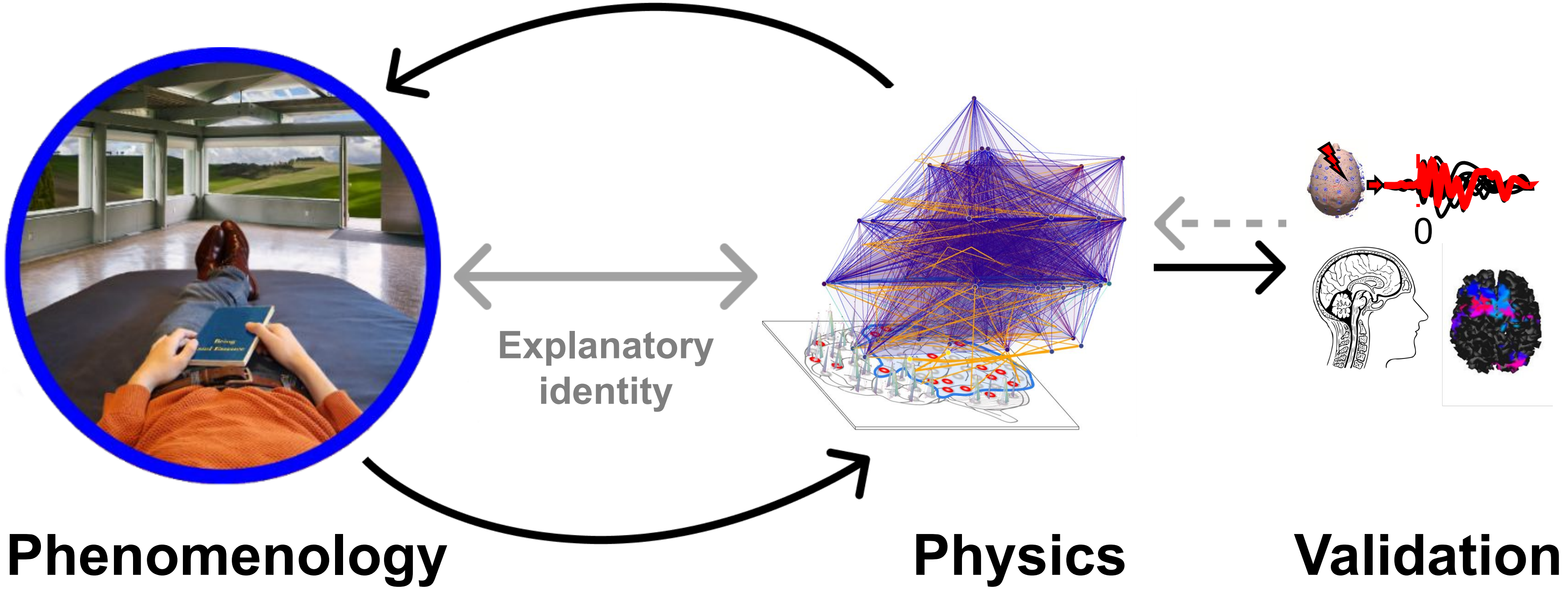
Phenomenology

Physics

How could the properties of my experience be accounted for by the properties of its physical substrate?

A Scientific Approach

Is this thing conscious? In what way?



Explanatory identity

Phenomenology

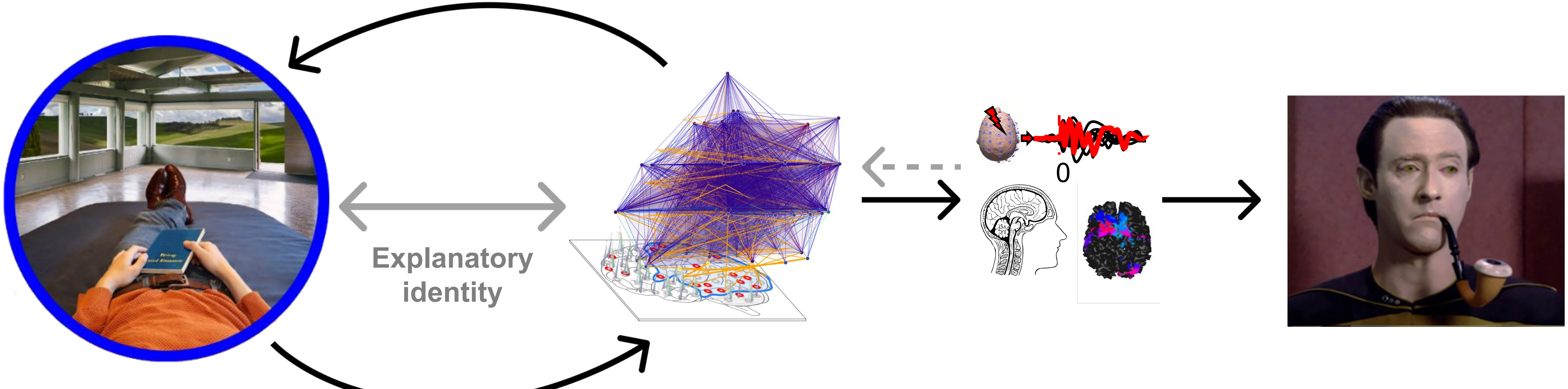
Physics

Validation

How could the properties of my experience be accounted for by the properties of its physical substrate?

A Scientific Approach

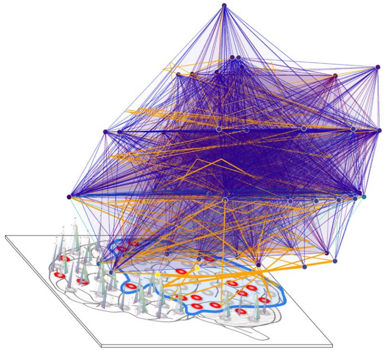
Is this thing conscious? In what way?



Phenomenology **Physics** **Validation** **Extrapolation**

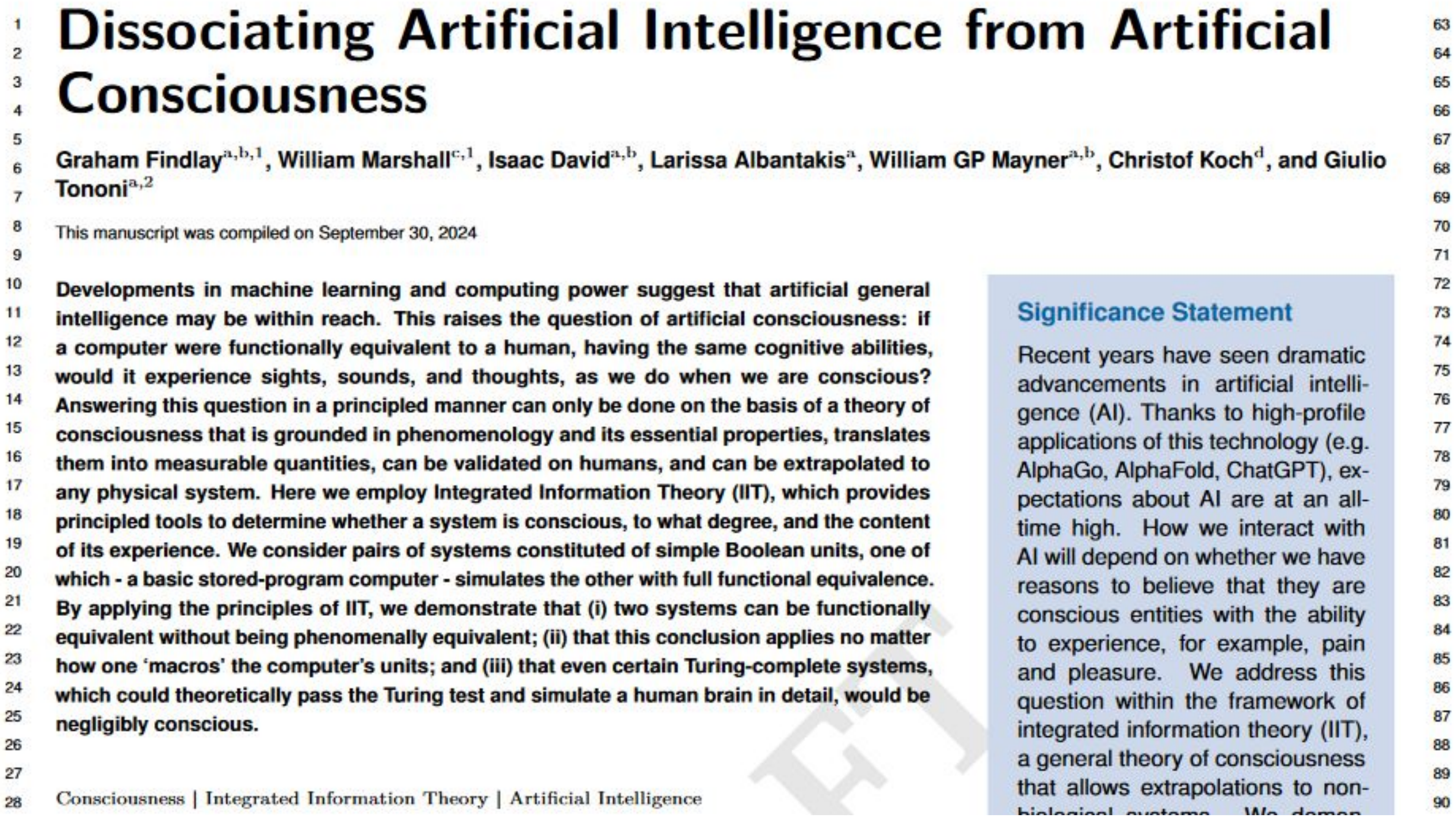
How could the properties of my experience be accounted for by the properties of its physical substrate?

Does intelligence imply consciousness?



?

Does intelligence imply consciousness?



Dissociating Artificial Intelligence from Artificial Consciousness

Graham Findlay[a,b,1], William Marshall[c,1], Isaac David[a,b], Larissa Albantakis[a], William GP Mayner[a,b], Christof Koch[d], and Giulio Tononi[a,2]

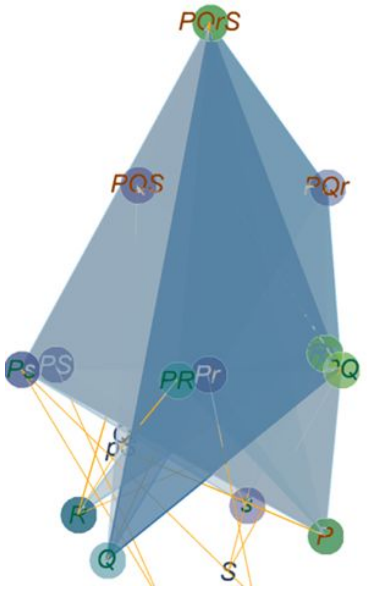
This manuscript was compiled on September 30, 2024

Developments in machine learning and computing power suggest that artificial general intelligence may be within reach. This raises the question of artificial consciousness: if a computer were functionally equivalent to a human, having the same cognitive abilities, would it experience sights, sounds, and thoughts, as we do when we are conscious? Answering this question in a principled manner can only be done on the basis of a theory of consciousness that is grounded in phenomenology and its essential properties, translates them into measurable quantities, can be validated on humans, and can be extrapolated to any physical system. Here we employ Integrated Information Theory (IIT), which provides principled tools to determine whether a system is conscious, to what degree, and the content of its experience. We consider pairs of systems constituted of simple Boolean units, one of which - a basic stored-program computer - simulates the other with full functional equivalence. By applying the principles of IIT, we demonstrate that (i) two systems can be functionally equivalent without being phenomenally equivalent; (ii) that this conclusion applies no matter how one 'macros' the computer's units; and (iii) that even certain Turing-complete systems, which could theoretically pass the Turing test and simulate a human brain in detail, would be negligibly conscious.

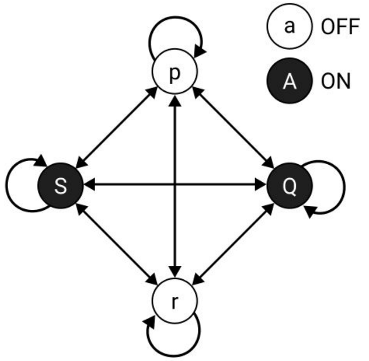
Consciousness | Integrated Information Theory | Artificial Intelligence

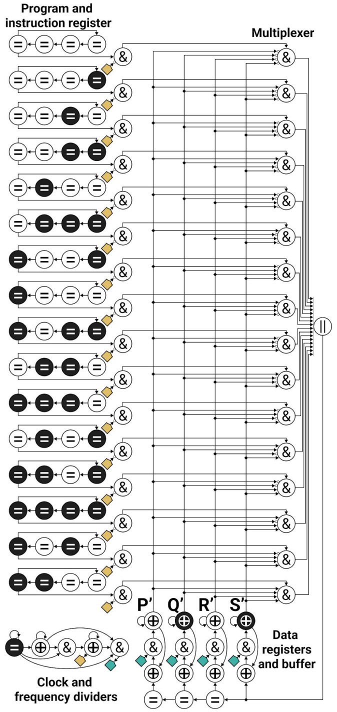
Significance Statement

Recent years have seen dramatic advancements in artificial intelligence (AI). Thanks to high-profile applications of this technology (e.g. AlphaGo, AlphaFold, ChatGPT), expectations about AI are at an all-time high. How we interact with AI will depend on whether we have reasons to believe that they are conscious entities with the ability to experience, for example, pain and pleasure. We address this question within the framework of integrated information theory (IIT), a general theory of consciousness that allows extrapolations to non-

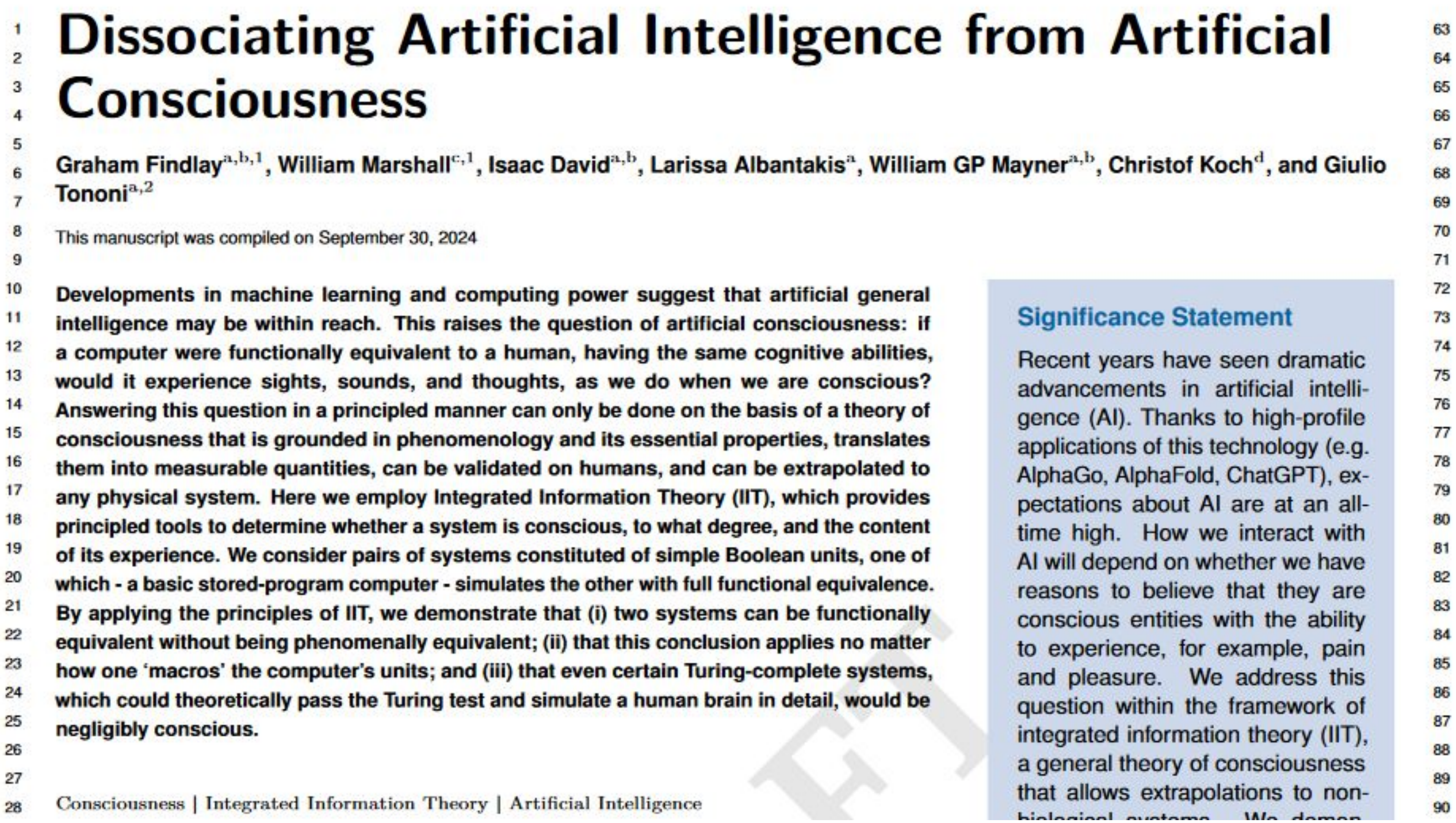


?




Program and instruction register
Multiplexer
P'
Q'
R'
S'
Data registers and buffer
Clock and frequency dividers

Does intelligence imply consciousness?



Dissociating Artificial Intelligence from Artificial Consciousness

Graham Findlay[a,b,1], William Marshall[c,1], Isaac David[a,b], Larissa Albantakis[a], William GP Mayner[a,b], Christof Koch[d], and Giulio Tononi[a,2]

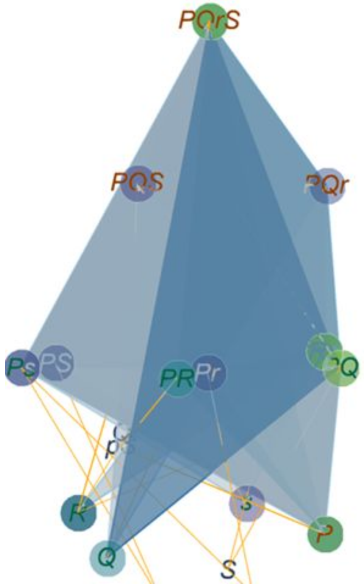
This manuscript was compiled on September 30, 2024

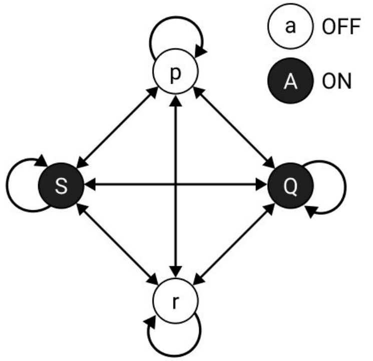
Developments in machine learning and computing power suggest that artificial general intelligence may be within reach. This raises the question of artificial consciousness: if a computer were functionally equivalent to a human, having the same cognitive abilities, would it experience sights, sounds, and thoughts, as we do when we are conscious? Answering this question in a principled manner can only be done on the basis of a theory of consciousness that is grounded in phenomenology and its essential properties, translates them into measurable quantities, can be validated on humans, and can be extrapolated to any physical system. Here we employ Integrated Information Theory (IIT), which provides principled tools to determine whether a system is conscious, to what degree, and the content of its experience. We consider pairs of systems constituted of simple Boolean units, one of which - a basic stored-program computer - simulates the other with full functional equivalence. By applying the principles of IIT, we demonstrate that (i) two systems can be functionally equivalent without being phenomenally equivalent; (ii) that this conclusion applies no matter how one 'macros' the computer's units; and (iii) that even certain Turing-complete systems, which could theoretically pass the Turing test and simulate a human brain in detail, would be negligibly conscious.

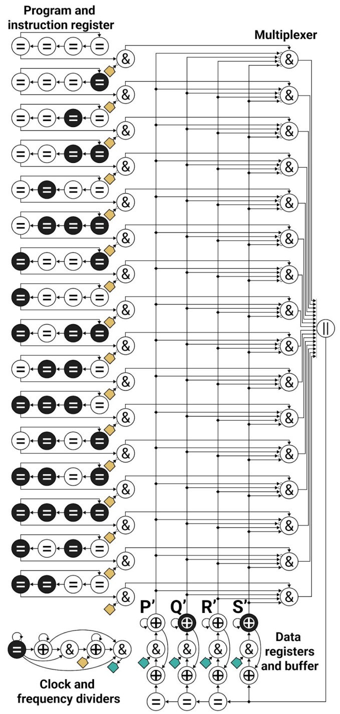
Consciousness | Integrated Information Theory | Artificial Intelligence

Significance Statement

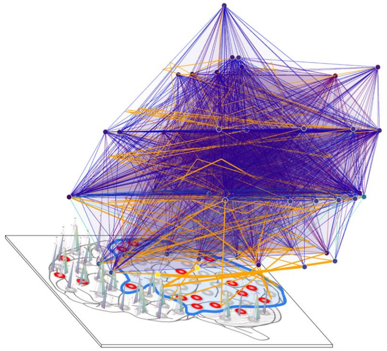
Recent years have seen dramatic advancements in artificial intelligence (AI). Thanks to high-profile applications of this technology (e.g. AlphaGo, AlphaFold, ChatGPT), expectations about AI are at an all-time high. How we interact with AI will depend on whether we have reasons to believe that they are conscious entities with the ability to experience, for example, pain and pleasure. We address this question within the framework of integrated information theory (IIT), a general theory of consciousness that allows extrapolations to non-biological systems. We demon






Program and instruction register
Multiplexer
P'
Q'
R'
S'
Data registers and buffer
Clock and frequency dividers

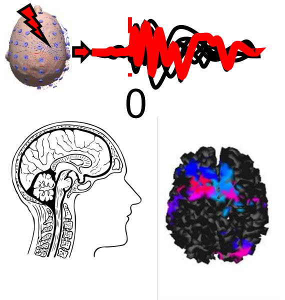
Does intelligence imply consciousness?



✗

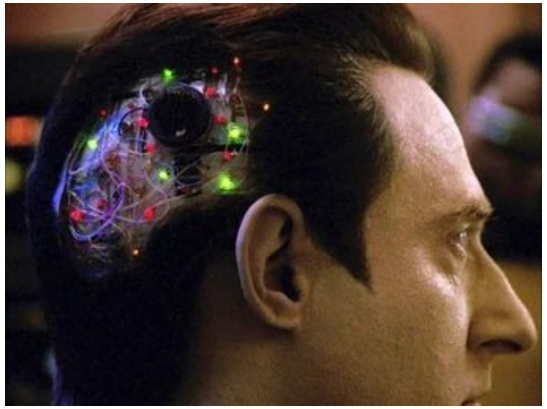
What comes next?

- More validation in IIT healthy adult humans
  - Develop computational and statistical tools to support empirical research



Validation

- Exploring alternative substrates for machine consciousness
  - Neuromorphic computers or quantum computers



Thanks for listening!

www.iit.wiki