Műszaki Tudományos Közlemények vol. 22. (2025) 59–64.
DOI: Hungarian: https://doi.org/10.33895/mtk-2025.22.10
English: https://doi.org/10.33894/mtk-2025.22.10

# INTERDISCIPLINARY AI RESEARCH METHODOLOGY

Levente KOLCSÁR,[1] László BAKÓ[2]

[1] *Sapientia University, Faculty of Technical and Human Sciences, Department of Electrical Engineering, Târgu Mureş, Romania, kolcsar.levente@ms.sapientia.ro*

[2] *University, Faculty of Technical and Human Sciences, Department of Electrical Engineering, Târgu Mureş, Romania, lbako@ms.sapientia.ro*

**Abstract**

The use of artificial intelligence (AI) in our time is becoming more and more extensive, so it is no question that there is a need for new AI concepts. We have developed our research methodology to accommodate new cognitive science theories, which we present through a concrete example: how Integrated Information Theory (IIT) was born and what it is good for. IIT is one of the main theories of consciousness in cognitive neuroscience and phenomenology. The main pillar of our methodology is the philosophy of science, which puts IIT into context, examines its origin and usability. This serves the purpose of a framework to be able to interpret ideas from the most diverse fields and to rank them according to their engineering applicability, so that we can create new, useful AI toolkits from them.

**Keywords***: AI, research, philosophy of science, IIT.*

## 1. Introduction

In our study, using the philosophy of science methodology, the AI researcher contextualizes theories from the field of cognitive sciences and only then begins working with them. This is necessary because interdisciplinary work requires order and a unified interpretation, which demands a well-defined framework. Through this method, the researcher becomes familiar with the theory itself, its origin, its mathematical description, and its applicability. We define the three main fields that constitute the structure of the methodology:

**α. Philosophy** consists of systematic linguistic and conceptual structures derived from our perceptions of the world. The central questions of philosophy are: What is nature? What is its origin? What is its foundation? Who are we? These ideas manifest not only in our understanding methods but also in our everyday lives.

**β. Science** as we define it, is a branching process that, like philosophy, seeks to understand nature. However, it grounds this understanding in systematic observation and experimentation.

This reveals that the overlap between the two fields occurs during the creation of new scientific theories, as in their earliest phase, devoid of experiments, researchers must rely solely on imagination. This connection, evident both logically and historicall [1], underscores that science originates from philosophy.

**γ. Philosophy of Science** is the branch of philosophy concerned with scientific thinking and methodology. Its primary aim is to systematically guide and refine scientific research and methods, while addressing the challenges of scientific reasoning and human bias. Examples of fields where we deem this discipline critical include: atomic physics, cosmology, biology, cognitive sciences, AI research, psychology, ethics, and the humanities. Conversely, we consider it less pivotal in areas such as: applied sciences, technologies, industry, and any domain where existing theories are applied unchanged.

## 2. Structure of the Methodology

The AI researcher must analyze and document the discovered or newly devised theory across the listed contexts, examining distinct aspects within each. The contexts correspond to the subjects defined by Greek letters (α, β, γ):

I. Historical Context (α, β)
II. Mathematical Description(α,β,γ)

III. Epistemological Context (β, γ)
IV. Ontological Context (α, β, γ)
V. Methodological Context (β, γ)
VI. Scientific Results (β)
VII. AI Applicability (β)
VIII. Axiological Context (α, γ)

For each context, the theory must be scored on a scale of 1–5, depending on its prominence in that context. A score of 1 indicates no relevance, while 5 denotes maximal relevance.

Example: The Integrated Information Theory (IIT) receives a score of 3 in the AI Applicability context. This scoring **Table 1** helps prioritize which theories warrant deeper engagement. The researcher must cite annotated sources in the documentation to ensure usability for future publications.

At the conclusion of the methodology, the theory's general properties are recorded, and a weighted composite score is calculated. The theory's AI applicability is then expressed as a percentage.

## 3. Methodology Example: IIT

We selected a representative theory that originated outside technical disciplines, exhibits strong context I. and II. relevance, and serves as a paradigmatic case for philosophical questions in science.

- Name: Integrated Information Theory (IIT); **[2]**
- Field: Cognitive Neuroscience and Phenomenology;
- Lead Author: Dr. Giulio Tononi (Neuroscientist);
- Theory: The causally integrated information within a network serves as a metric for determining its potential "consciousness". Denoted as Φ, it quantifies the extent to which a network's structure integrates causal information in itself.

**Table 1.** *Score table for IIT*

| Context | Points |
|---|---|
| I. | 5 |
| II. | 5 |
| III. | 3 |
| IV. | 1 |
| V. | 4 |
| VI. | 4 |
| VII. | 3 |
| VIII. | 4 |

### 3.1. Historical Context (I.)

In this context, we examine the pre-theoretical intellectual traditions or scientific procedures that preceded the theory historically, as well as its relationship to classical philosophical movements. While this section can often be omitted, it is recommended for theories with deep historical roots. IIT scores 5 here, as it originates from a longstanding philosophical problem: What is consciousness? Why is it so difficult to explain? Dr. Tononi provides a robust historical and literary overview in his book Phi **[3]**, where he explains how thinkers like Galileo grappled with questions of consciousness.

The "hard problem of consciousness" was famously articulated by philosopher David Chalmers in the modern era **[4]**: How do biological mechanisms give rise to unified, qualitative subjective experience? Why does red appear as "red"? Why is consciousness so specific and qualitative?

From this problem IIT constructed its main axioms. (**Figure 1**)

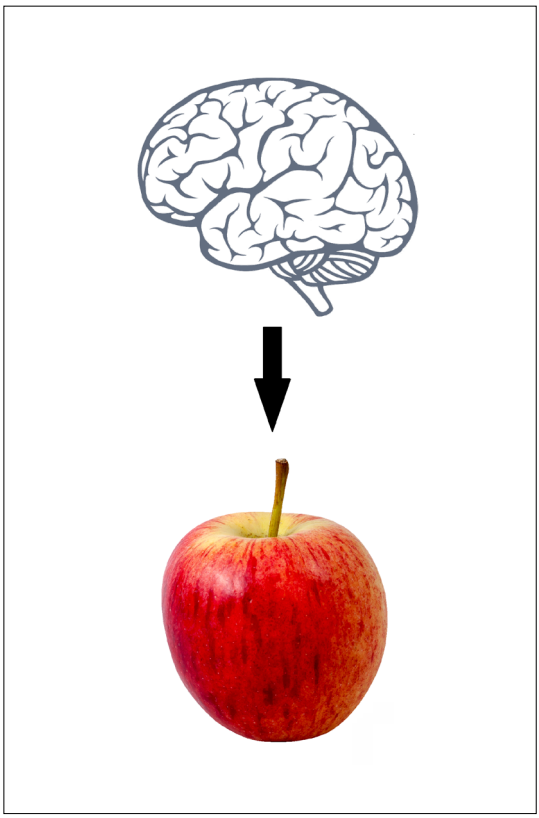

**Fig. 1.** *The "hard problem of consciousness"* **[5]**

### 3.2. Mathematical Description (II.)

In this section, we study the formal mathematical description of the theory, including all axioms, equations, and relationships that form its foundation. Axioms are properties derived from general observations, postulates are theoretical statements of derived regularities, and mathematical expressions formalize these postulates. In this context IIT scores 5 points.

#### 3.2.1. Existence

Axiom: Consciousness exists; therefore, it must possess causal power.

Postulate: The network must exhibit cause-effect power.

Mathematical Expression: Given a stochastic network in a state $x_t$ with transition probability matrix $T$, describing the probabilities of transitions from $x_t$ to $x_{t+1}$. From this we define the network's cause-effect repertoire:

$$P_{cause}(x_{t-1} \mid x_t)$$
$$P_{effect}(x_{t+1} \mid x_t) \qquad (1)$$

#### 3.2.2. Composition

Axiom: Consciousness is composed of distinguishable phenomenological elements (e.g., sounds, colors, sensations).

Postulate: The network must form a complex of mechanisms (subsets of elements) that collectively constrain the cause-effect repertoire.

Mathematical Expression:

For a fully connected 3-node network:

$$M = \{A, B, C, AB, AC, BC, ABC\} \qquad (2)$$

where $M$ is the set of mechanisms. A mechanism qualifies as a complex if its calculated $\Phi > 0$. (Mechanisms include bidirectional connections, e.g. $AB = \{AB, BA\}$)

#### 3.2.3. Information

Axiom: Every conscious experience is specific and distinct from other possible conscious states.

Postulate: The network must exhibit a differentiated causal structure (high informational divergence from randomness).

Mathematical Expression:

$$EI_{cause} = D(P_{cause}, P_{max})$$
$$EI_{effect} = D(P_{effect}, P_{max})$$
$$EI = min(EI_{cause}, EI_{effect}), \qquad (3)$$

where:

$EI_{cause}$ – cause information;
$EI_{effect}$ – effect information;
$P_{max}$ – maximum entropy distribution;
$EI$ – effective information;
$D$ – distance metric (e.g., Earth Mover's Distance or Kullback–Leibler Divergence)

#### 3.2.4. Integration

Axiom: Consciousness is unified, not reducible to independent parts.

Postulate: integrated information (Φ) represents the irreducibility of information in a mechanism.

Mathematical term: bipartitions of the network: $\{\{A\}, \{B,C\}\}, \{\{B\}, \{A,C\}\}, \{\{C\}, \{A,B\}\}\}$

$$\Phi = min_{MIP}[D(P_{total}, P_{partitioned})] \qquad (4),$$

where:

$\varphi > 0$, if the subsystem is irreducible;
$P_{total}$ – cause-effect repertoire of the network;
$P_{partitioned}$ – cause-effect repertoire of the partitioned network;
$MIP$ – minimum information partition.

#### 3.2.5. Exclusion

Axiom: Consciousness is a specific state in time and space, excluding other states.

Postulate: The network's major complex is the one with maximal Φ.

Mathematical Expression:

$$\Phi_{max} = arg\ max_{m \in M}\{\Phi(m)\} \qquad (5)$$

### 3.3. Epistemological Context (III.)

This section evaluates how the theory expands current knowledge and its explanatory power, a critical criterion for applicability. In AI, we aim to implement as many useful properties as possible to develop better problem-solving artificial agents.

Natural Agent (Living Being) → Artificial Agent (AI)

Properties of natural agents are extrapolated to artificial agents. Cognitive science studies the cognitive properties of human beings. IIT quantifies consciousness through the metric Φ. For this section, the theory scores an average of 3 points.

### 3.4. Ontological Context (IV.)

This context examines the ontological framework (e.g., electromagnetism, cells, atoms, or unconventional paradigms) the theory uses in its explanations.

IIT is ontologically independent, akin to Shannon's information theory, meaning it applies to any medium since it focuses on state dynamics rather than the specific processes or elements

constituting those states. For example, IIT can model both neurons and logic gates. This ontological independence is valuable for AI research, as such theories can be adapted to technical domains. IIT scores 1 point here.

### 3.5. Methodological Context (V.)

This section assesses the extent to which the theory aligns with established scientific methodologies:

Deductive Method: 0. Background → 1. Theory → 2. Prediction Method → 3. Testing the Prediction (Measurement) → 4. Validate the Prediction → 5. Accept/Reject the Theory

Inductive Method: 0. Background → 1. Observation (Measurement) → 2. Identify Patterns → 3. Generalize → 3. Theory → 4. Validate the Theory → 5. Accept/Reject the Theory.

IIT follows a deductive methodology. While it has demonstrated neuroscientific results, its computational complexity (NP-hard) for large networks necessitates approximations. Thus, it scores 4 points in this context.

### 3.6. Scientific Results (VI.)

Here, we evaluate how well the theory is validated through observational and experimental methods. A high score reflects robust explanatory power for natural phenomena.

IIT scores 4 points in this context because it has produced predictive results comparable to the Global Workspace Theory of consciousness [6]. Additional achievements in brain research include. [7, 8] (Note: A score of 1 is assigned if no experimental results exist. Findings within the theory's own field suffice for validation.)

### 3.7. AI Applicability (VII.)

For the IIT theory, a public Python library (pyphi) has been developed, freely available under the condition that associated publications are properly cited. [9] This library enables the calculation of Φ for arbitrary n-node stochastic networks. Figure 2 illustrates our example, in which we compute Φ for a network of logic gates.

Python code:

```
#The name of the nodes and it's order:
  node_labels = ('OR','AND','XOR' )
#Initial states vector [n]:
  state = (1, 0, 0)
#Connectivity matrix [n * n]:
  cm = np.array ([[0,1,1],
                  [1,0,1],
                  [1,1,0]])
#Transition probability matrix [2^n * 2^n]:
  tpm = np.array([[1, 0, 0, 0, 0, 0, 0, 0 ],
                  [0, 0, 0, 0, 1, 0, 0, 0 ],
                  [0, 0, 0, 0, 0, 1, 0, 0 ],
                  [0, 1, 0, 0, 0, 0, 0, 0 ],
                  [0, 1, 0, 0, 0, 0, 0, 0 ],
                  [0, 0, 0, 0, 0, 0, 0, 1 ],
                  [0, 0, 0, 0, 0, 1, 0, 0 ],
                  [0, 0, 0, 1, 0, 0, 0, 0 ]])
#Note:
cm connectivity entries can take binary values (0 or 1). In this example, the main diagonal is set to 0 because the network does not include nodes with self-connections.
tpm is a matrix that encodes the probabilities of transitioning from all possible states at t (row indices) to all possible state at t+1
#Generating the network:
  net = pyphi.Network(tpm, cm, node_labels)
#Compute the major complex of the network:
  mct = pyphi.compute.major_complex(net, state)
#Use Φ of the major complex:
  phi_maxt = mc.phi
#Φmaxt = 1.916
```

Currently, we can state that IIT is useful for network analysis to determine the degree of recurrence (feedback loops) in a network, as feedback increases the Φ value. According to the model, the more recurrent pathways a network has, the more integrated it becomes. This property can also be interpreted as dynamic memory, determined not only by the states of nodes but also by

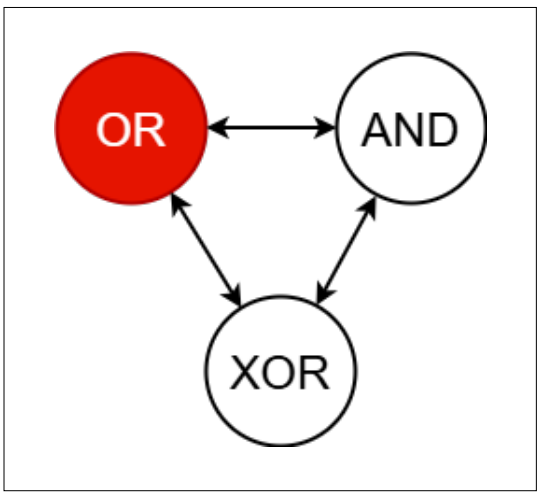


**Fig. 2.** *Example network (red "1", white "0").*

the network's topology. While this is a network analysis tool, we do not yet see direct utility for AI, so we assign a neutral score of 3 for AI applicability. Potential applications include using Φ as a training criterion for networks with variable topologies or, in the future, as a metric for ethical classification of AI systems. This context requires further analysis, as we believe Φ could also be applied to network optimization and sensor fusion problems.

### 3.8. Axiological Context (VIII.)

This section evaluates the theory's relevance to AI ethics. AI ethics is an increasingly critical issue, making this context significant for many models. According to the nonprofit Eleos AI **[10]** two main risks threaten AI evolution:

– Overestimation (attributing human-like properties to AI);
– Underestimation.

IIT scores 4 here, as its Φ metric suggests even logic gates could form a "conscious" AI. However, we remain critical of this claim. While Φ is currently treated as a metric—not a definitive measure—we acknowledge its potential role in understanding consciousness. We intentionally avoid defining consciousness or intelligence, as these remain open questions in neuroscience. For AI, we aim to maintain flexibility to develop new models. IIT posits consciousness as an internal process distinct from intelligence (an output property.

### 3.9. Methodology Summary

At the end of the documentation, we record the general properties of the IIT theory:

– Brain inspired AI? Yes;
– Traditional/Statistical AI? No;
– Top-down model? Only in method;
– Bottom-up model? No;
– Deductive or inductive? Inductive.

Using the context scores (**Table 1**) we calculate a weighted composite score, prioritizing AI applicability (VII.) while heavily weighting mathematical rigor (II.) and scientific results (VI.). Contexts III., IV., V., and VIII. contribute indirectly, while I. (historical) is excluded. Context IV. is inversely weighted, as stronger ontological foundations complicate interdisciplinary adaptation. This weighting requires further refinement.

$$\text{sum} = \text{I.}*0 + \text{II.}*0.5 + \text{III.}*0.25 - \text{IV.}*0.25 + \text{V.}*0.25 + \text{VI.}*0.5 + \text{VII.}*1 + \text{VIII.}*0.1$$

Converted to a percentage using the rule of three:

$$IIT_{rating} = 74\%.$$

## 4. Conclusions

Our methodology is a starting point for deriving practical AI tools from cognitive science theories. While it requires refinement through application to more theories, the goal is to enable analysis of any theory or model, regardless of its cognitive science origin. A repository of such analyses would simplify designing new models by leveraging existing insights and identifying promising applications for implementation.

The representative IIT theory discussed by the methodology requires further analysis on our part in terms of its applicability in AI, which will be a separate task for us in the near future.

## References


[1] Quack M.: *Science and Arts, Philosophy and Science: Why after All? Why Not?*, Helvetica Chimica Acta, 106/4. (2023) 1–3. https://doi.org/10.1002/hlca.202200174

[2] Oizumi M., Albantakis L., Tononi G.: *From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0*. PLOS Computational Biology, 2014, 1–25. https://doi.org/10.1371/journal.pcbi.1003588

[3] G.: *Phi: A Voyage from the Brain to the Soul.* 1. ed. Pantheon Books, New York, 2012, 1–38.

[4] Chalmers D.: *Facing up to the Problem of Consciousness*. Journal of Consciousness Studies, 2/3. (1995) 200–219. https://philpapers.org/rec/CHAFUT

[5] UI Here free samples. https://www.uihere.com/ (2025.02.22)

[6] Cogitate Consortium, Ferrante O., Gorska-Klimowsk U., et al.: *An Adversarial Collaboration to Critically Evaluate Theories of Consciousness.* bioRxiv 2023.06.23.546249, 1–68. https://doi.org/10.1101/2023.06.23.546249

[7] Ferrarelli F., Massimini M., Sarasso S., Casali A., Riedner B.A., Angelini G., Tononi G., Pearce R.A.: *Breakdown in Cortical Effective Connectivity during Midazolam-Induced Loss of Consciousness*. Proc Natl Acad Sci USA. 107/6. (2010) 2681–2686. https://doi.org/10.1073/pnas.0913008107

[8] Casali A.G., Gosseries O.: *A Theoretically Based Index of Consciousness Independent of Sensory Processing and Behavior*. Science Translational Medicine, 5. (2013) 1–12. https://hdl.handle.net/2268/171542

[9] Mayner W.G.P., Marshall W., Albantakis L., Findlay G., Marchman R., Tononi G.: *PyPhi: A Toolbox for Integrated Information Theory.* PLoS Computational Biology, 14/7. (2018) 1–21. https://doi.org/10.1371/journal.pcbi.1006343

[10] Long R., Sebo J., Butlin P., Finlinson K., Fish K., Harding J., Pfau J., Sims T., Birch J., Chalmers D.: *Taking AI Welfare Seriously*. arXiv:2411.00986, (2024) 1–3.
https://doi.org/10.48550/arXiv.2411.00986