Consciousness & AI

15 June 2023

Ziyuan Ye

Overview

- What is consciousness?
  I. Phenomenal Consciousness
  II. Cognitive (Access) Consciousness

- Where does the consciousness come from?
  I. Integrated Information Theory
  II. Global Neuronal Workspace Theory
  III. Dendrites Integration Theory

- Discussion about consciousness and next-generation AI models.

Overview

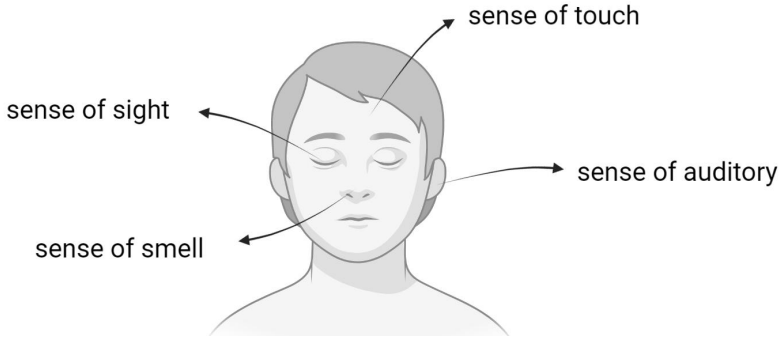
- What is consciousness?
  I. Phenomenal Consciousness
  II. Cognitive (Access) Consciousness

- Where does the consciousness come from?
  I. Integrated Information Theory
  II. Global Neuronal Workspace Theory
  III. Dendrites Integration Theory

- Discussion about consciousness and next-generation AI models.

What is consciousness?

Phenomenal consciousness: subjective experience



An analogue mapping between the nervous system and patterns of information in the environment or in the body.

Highly related to our perception!

Cognitive (Access) consciousness: cognition and processing of certain information

Cognitive Ability: attention, memory, reasoning, problem-solving, language processing…
Knowledge and Learning: knowledge acquisition, processing, storage, application…
Emotional Cognition: understanding and recognition of our own emotions and those of others
Self-awareness: understanding and perception of ourselves (e.g., self-concept)
…

Block, N. (2005). Two neural correlates of consciousness. *Trends in cognitive sciences*, *9*(2), 46-52.

Overview

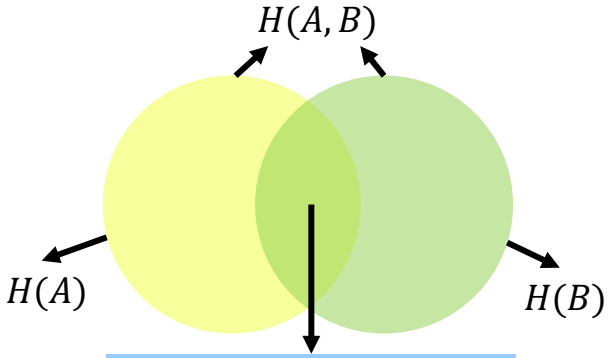
- What is consciousness?
  I. Phenomenal Consciousness
  II. Cognitive (Access) Consciousness

- Where does the consciousness come from?
  I. Integrated Information Theory
  II. Global Neuronal Workspace Theory
  III. Dendrites Integration Theory

- Discussion about consciousness and next-generation AI models.

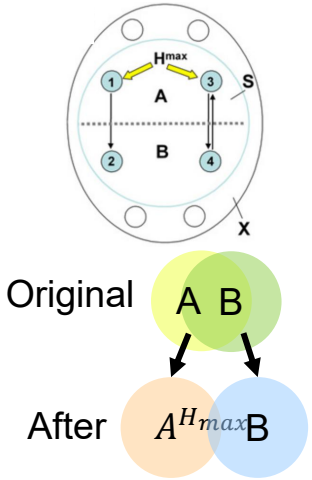
Integrated Information Theory (IIT)

Consciousness corresponds to the capacity of a system to integrate information. The state of consciousness can be measured as the causally effective information (EI) ϕ **value** of a complex of elements.



Mutual information(MI)

$$MI(A;B) = H(A) + H(B) - H(AB)$$



Effective information from A to B

$$EI(A \rightarrow B) = MI(A^{H_{max}}; B)$$

Effective information between A and B

$$EI(A \leftrightarrow B) = MI(A^{H_{max}}; B) + MI(\mathrm{A}; \mathrm{B}^{H_{max}})$$

Normalized effective information between A and B

$$EI(\widetilde{A \leftrightarrow B}) = EI(A \leftrightarrow B) \,/\, H^{max}(A \leftrightarrow B)$$

$$H^{max}(A \leftrightarrow B) = \min(A^{H_{max}}; B^{H_{max}})$$

Minimum information bipartition(MIB)

$$^{MIB}A \leftrightarrow B = argmin\{EI(\widetilde{A \leftrightarrow B})\}$$

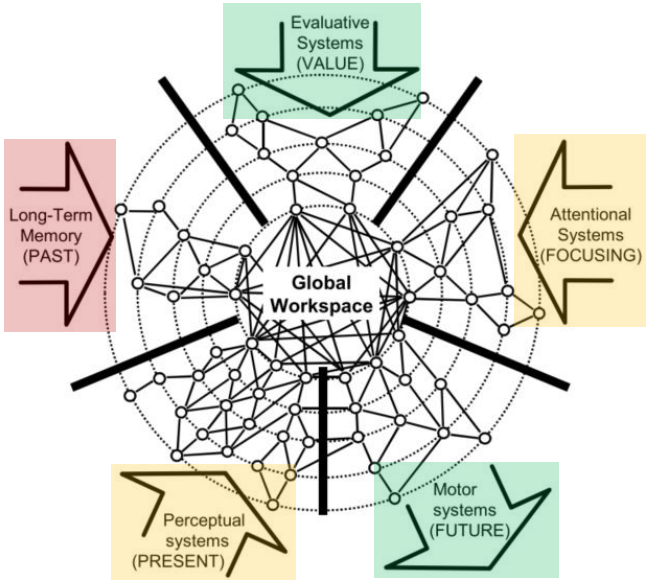
Information integration of subset S

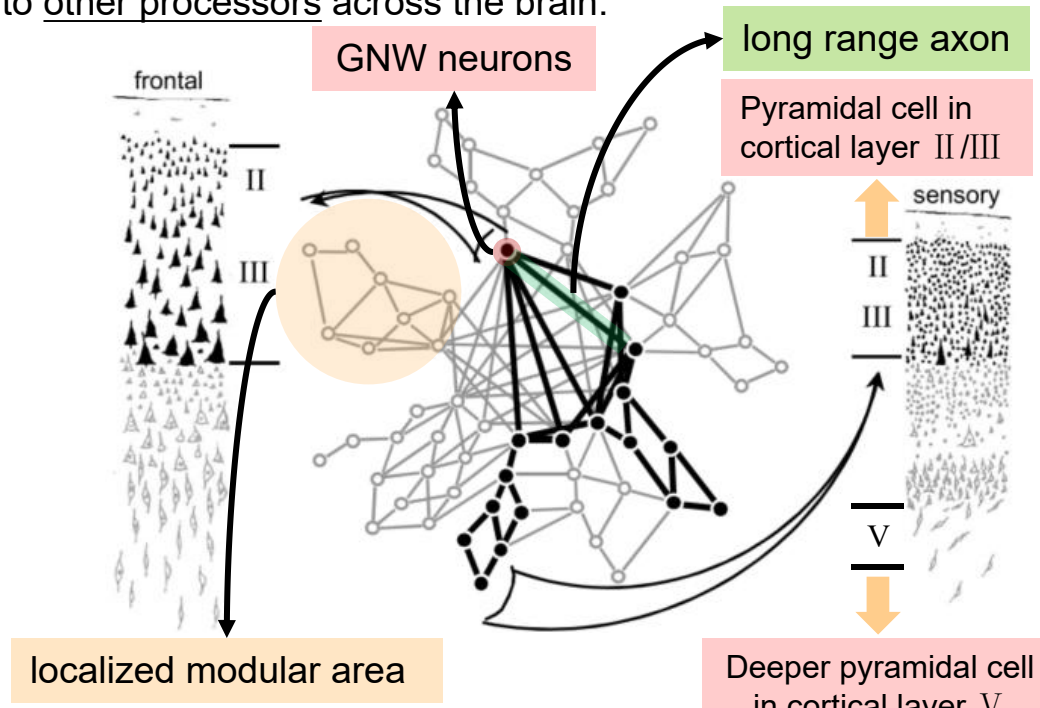
$$\phi(\mathrm{S}) = \mathrm{EI}(\,^{MIB}A \leftrightarrow B)$$

Tononi, G. (2004). An information integration theory of consciousness. *BMC neuroscience*, *5*, 1-22.

Global Neuronal Workspace Theory (GNWT)

Perceptual contents are acted upon by localized processors, only become conscious when they are widely broadcasted to other processors across the brain.

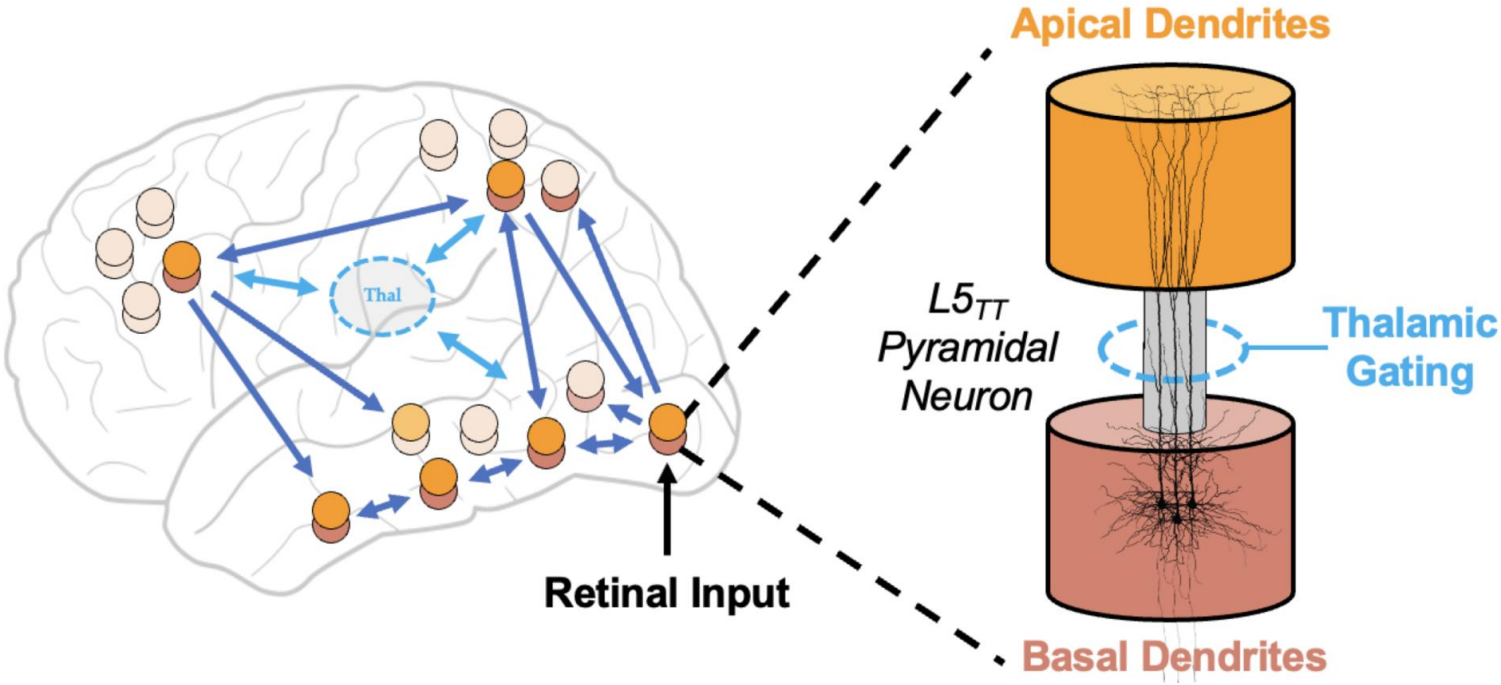




Mashour, G. A., Roelfsema, P., Changeux, J. P., & Dehaene, S. (2020). Conscious processing and the global neuronal workspace hypothesis. *Neuron*, *105*(5), 776-798.

Dendrites Integration Theory (DIT)

Conscious states and content emerge in deep pyramidal neurons.



Aru, J., Larkum, M., & Shine, J. M. (2023). The feasibility of artificial consciousness through the lens of neuroscience. *arXiv preprint arXiv:2306.00915*.

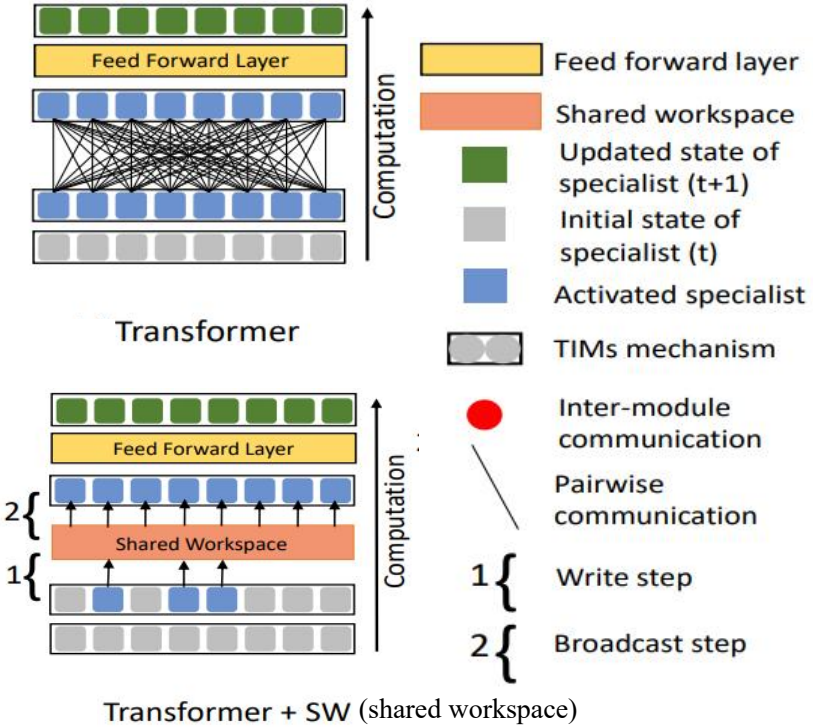
Overview

- **What is consciousness?**
  - I. Phenomenal Consciousness
  - II. Cognitive (Access) Consciousness

- **Where does the consciousness come from?**
  - I. Integrated Information Theory
  - II. Global Neuronal Workspace Theory
  - III. Dendrites Integration Theory

- **Discussion about consciousness and next-generation AI models.**

Transformer + GNWT



Two step process

Write step: write information into workspace

Embedding: $\boldsymbol{R} \in \mathbb{R}^{n_s \times n_h}$ (specialist, hidden state)

Each row of $\boldsymbol{R}$: $h_t^k \in \mathbb{R}^{1 \times n_h}$, $k \in \{1, \ldots, n_s\}$, t: stage number

Shared workspace: $\boldsymbol{M} \in \mathbb{R}^{1 \times n_h}$

Query: $\widetilde{\boldsymbol{Q}} = \boldsymbol{M}\widetilde{\boldsymbol{W}}^{\boldsymbol{q}}$ **Key/Value:** $\boldsymbol{R}$

Update Shared Workspace: $M \leftarrow softmax\left(\frac{\widetilde{\boldsymbol{Q}}(\boldsymbol{R}\widetilde{\boldsymbol{W}}^{\boldsymbol{e}})^T}{\sqrt{d_e}}\right)\boldsymbol{R}\widetilde{\boldsymbol{W}}^{\boldsymbol{v}}$

Soft competition

Hard competition: Select Top-k specialists

Broadcast step: broadcast information from workspace

Query: $\widehat{\boldsymbol{q_k}} = h_t^k \widehat{\boldsymbol{W}}^{\boldsymbol{q}}$

Key: $\widehat{\kappa_j} = (\boldsymbol{m_j}\widehat{\boldsymbol{W}}^{\boldsymbol{e}})^{\mathrm{T}}$

Value: $\widehat{\boldsymbol{v_j}} = \boldsymbol{m_j}\widehat{\boldsymbol{W}}^{\boldsymbol{v}}$

$$\boldsymbol{h_t^k} \leftarrow \boldsymbol{h_t^k} + \sum_j \mathrm{softmax}\left(\frac{\widehat{\boldsymbol{q_k}}\widehat{\kappa_j}}{\sqrt{d_e}}\right)\widehat{\boldsymbol{v_j}}$$

Goyal, A., Didolkar, A. R., Lamb, A., Badola, K., Ke, N. R., Rahaman, N., ... & Bengio, Y. (2022) Coordination Among Neural Modules Through a Shared Global Workspace. In *International Conference on Learning Representations* (Oral).

Discussion

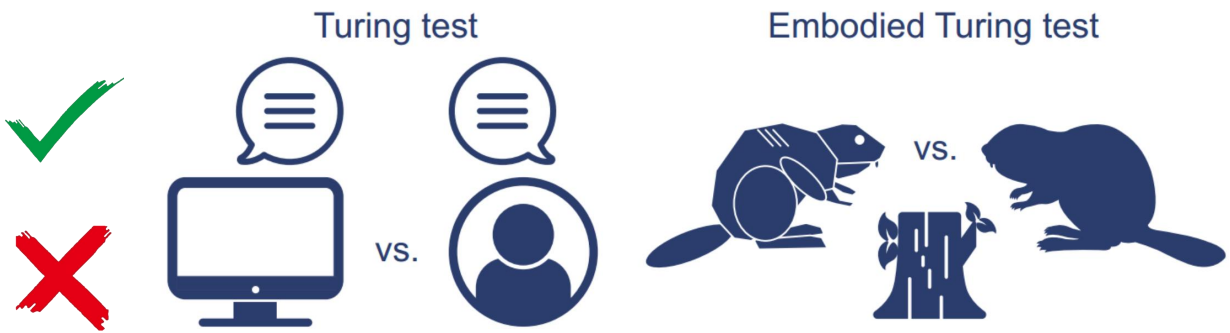
1. Measure a AI model's consciousness by using Integrated Information Theory. Improve model consciousness by optimizing model architecture design.

2. Design novel AI model architecture with Global Neuronal Workspace Theory / Dendrites Integration Theory.

How can we go beyond LLM?

Cognitive (Access) consciousness: cognition and processing of certain information

Phenomenal consciousness: subjective experience



Zador, A., Escola, S., Richards, B., Ölveczky, B., Bengio, Y., Boahen, K., ... & Tsao, D. (2023). Catalyzing next-generation artificial intelligence through neuroai. *Nature communications*, *14*(1), 1597.

Thanks for your attention!