What is consciousness, and could machines have it?

Psy 3280 – Week 10 Lecture (1 October 2018)

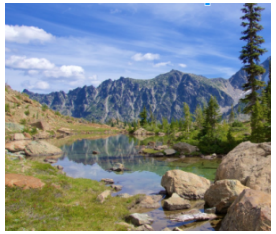
Rafik Hadfi
Zhao Hui Koh

Learning objective

- Can machines have consciousness?
- The multiple meanings of consciousness
  - C0: Unconscious processing
  - C1: Global availability of information
  - C2: Self monitoring
- Relationships between C1 and C2
- Pathways to artificial consciousness
  - Adversarial learning (Dehaene)
  - Maximizing Information Integration (IIT)
  - Minimizing Prediction Error (Predictive coding)

From human to artificial perception

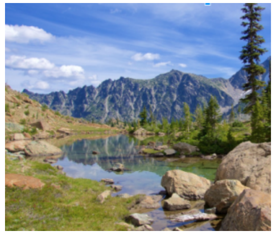
- The natural scene experiment is an example of how perception or report could be artificially replicated

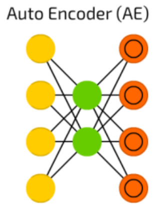


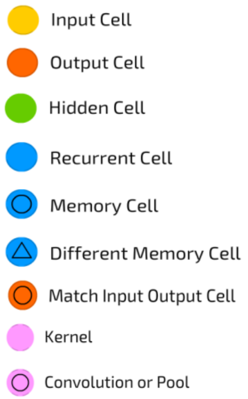
From human to artificial perception

- The natural scene experiment is an example of how perception or report could be artificially replicated using Autoencoders

Perception

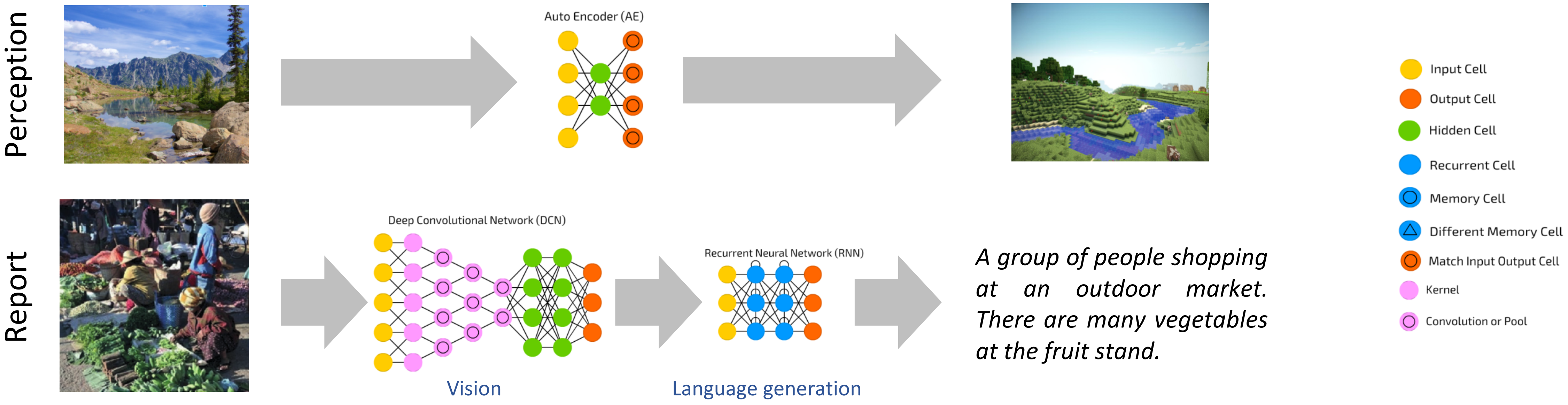






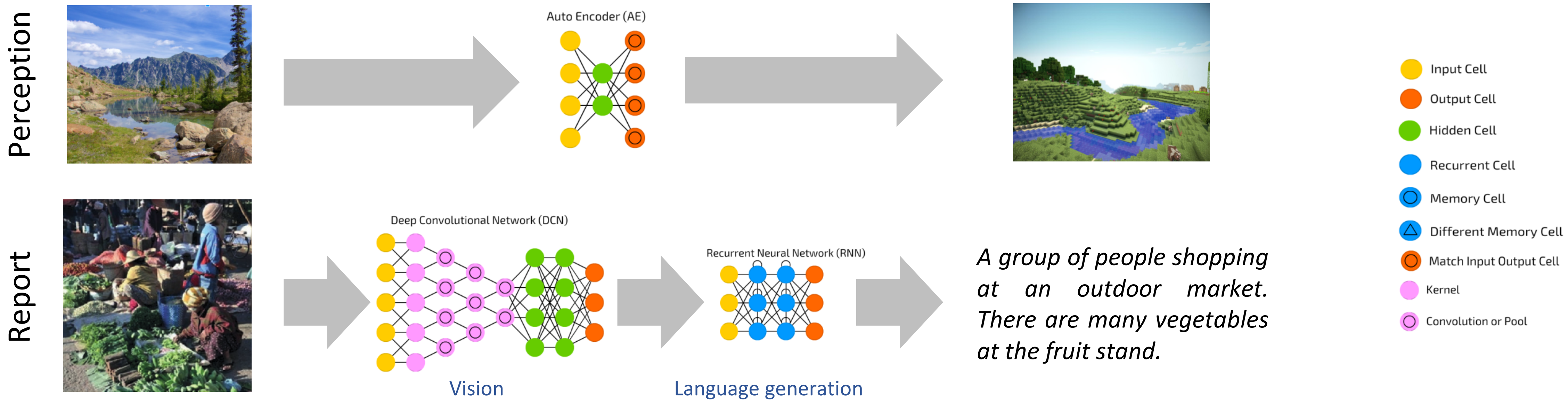
From human to artificial perception

- The natural scene experiment is an example of how perception or report could be artificially replicated using Autoencoders or Convolutional Nets



From human to artificial perception

- The natural scene experiment is an example of how perception or report could be artificially replicated using Autoencoders or Convolutional Nets



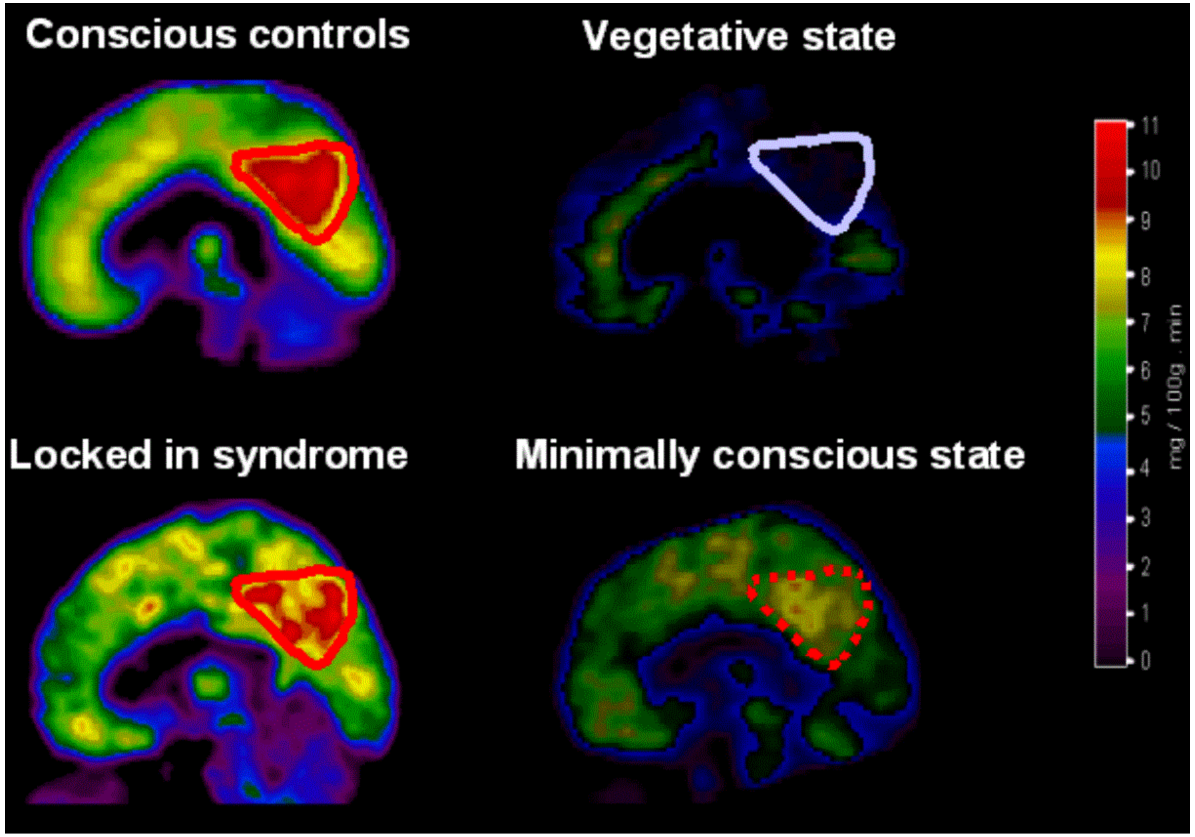
- What about consciousness, can machines have it?

Can machines have consciousness?

- To answer the question we must carefully consider how consciousness arises in the only physical system that undoubtedly possesses it: the human brain
- Neuroscientists have developed tools and theories to understand consciousness in the human brain:

Can machines have consciousness?

- Brain imaging



Can machines have consciousness?

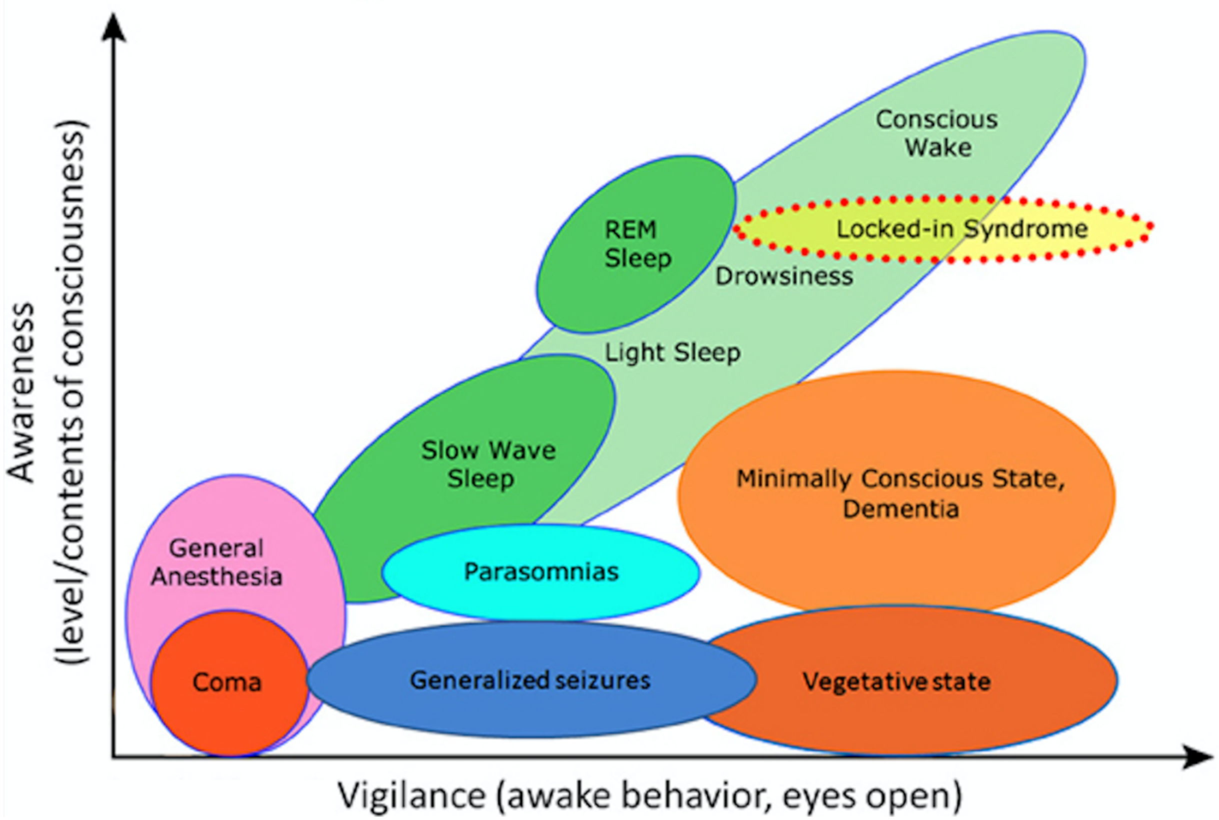
- Psychophysical experimentation

Table 1. Relative strengths of various psychophysical techniques for erasing a stimulus from visual awareness

[a]Variety of stimuli – is the technique effective at rendering a wide variety of stimuli invisible? [b]Stimulus size – does the technique work over a wide range of stimulus sizes? [c]Visual field location – does the technique work equally well in central and in peripheral vision? [d]Temporal aspects of stimulation – are there constraints on the exposure duration or on the timing of the stimulus? [e]Unambiguous invisibility – does the state of unawareness involve complete, unambiguous invisibility of the stimulus? [f]Invariant stimulation – does physical stimulation remain invariant when visual awareness fluctuates? [g]Duration – do the periods of unawareness last for longer than a few hundred milliseconds? [h]Predictability – is the onset of unawareness controllable, and are the durations of unawareness predictable?

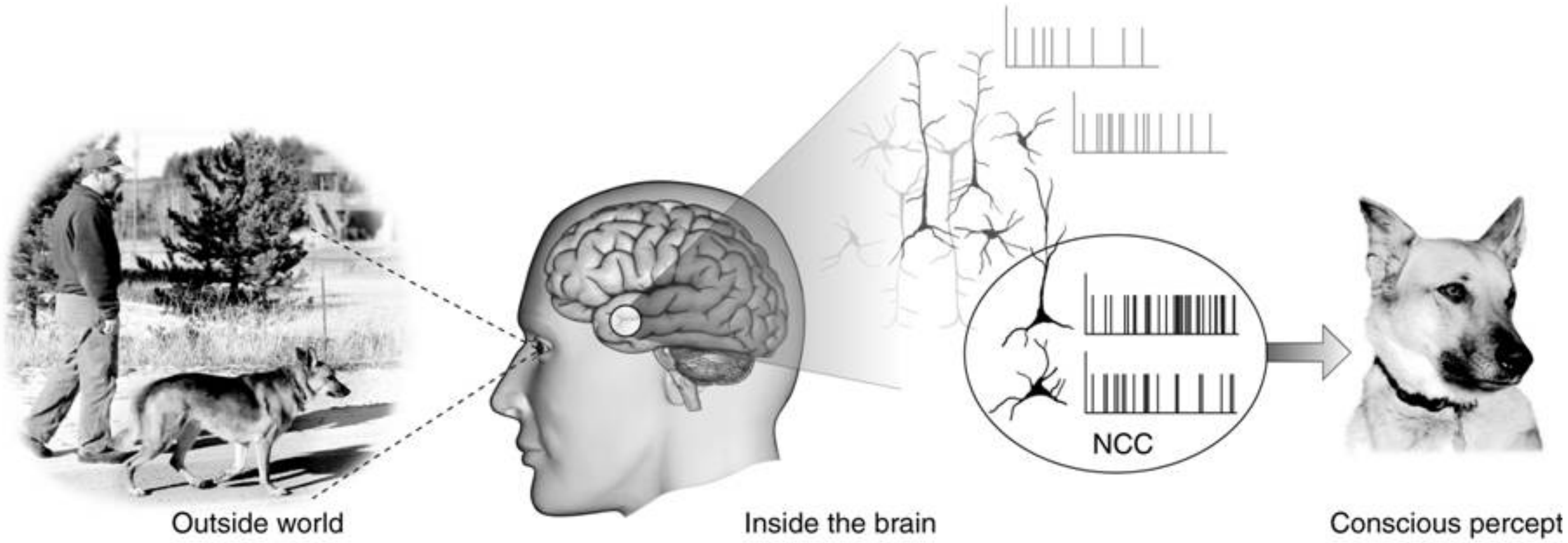
Can machines have consciousness?

- Drawing distinctions between the content and levels of consciousness



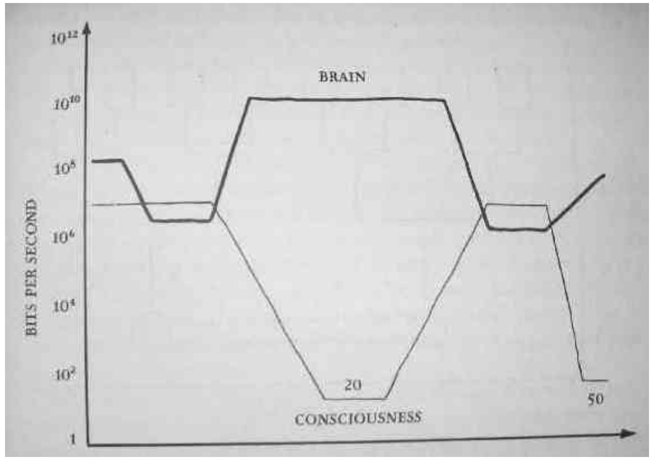
Can machines have consciousness?

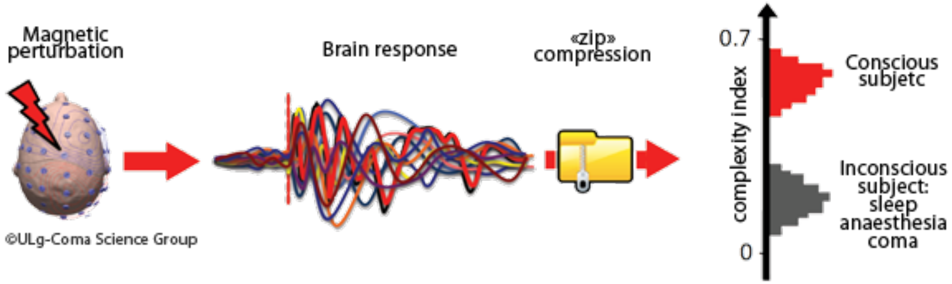
- Studying the neural correlates of consciousness (NCC)

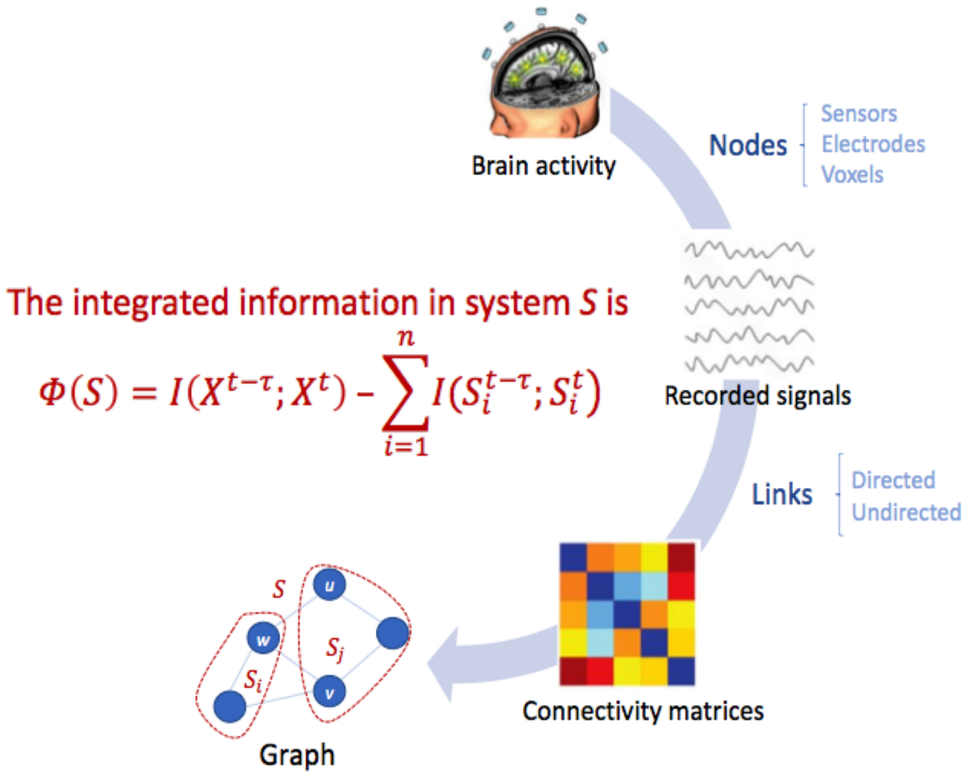


Can machines have consciousness?

- Quantifying consciousness: Bandwidth of Consciousness (BoC), Perturbational Complexity Index (PCI) or Integrated Information (Phi)

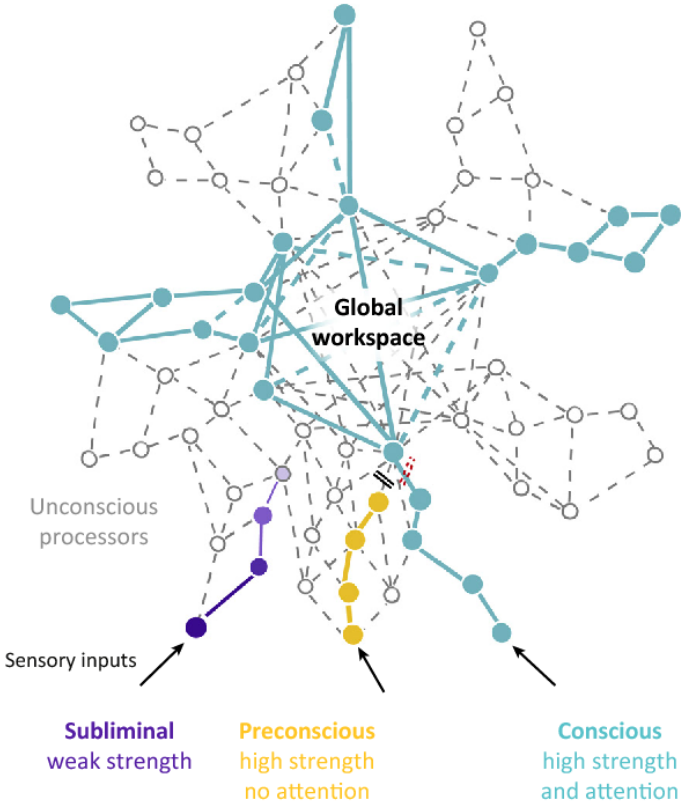


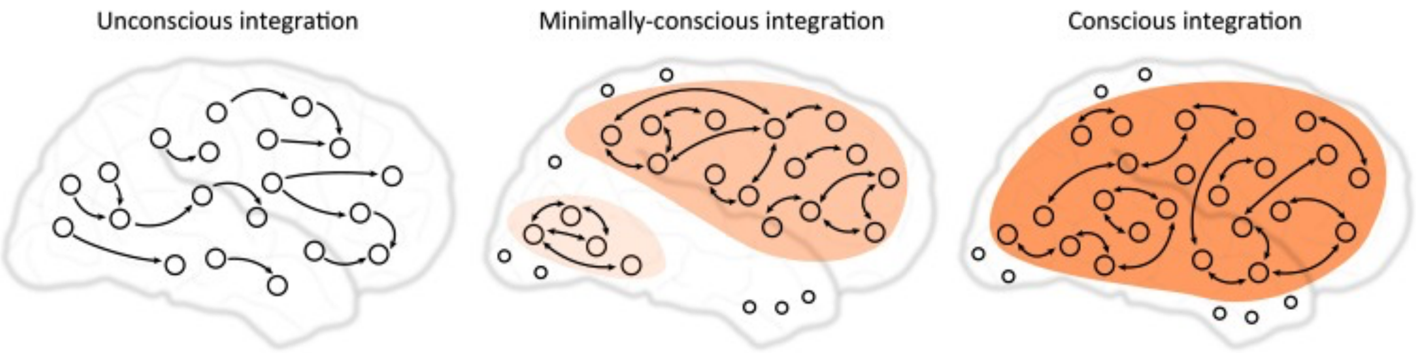




Can machines have consciousness?

- Proposing theories of consciousness: Global workspace Theory, Integrated Information Theory





The multiple meanings of Consciousness

- Let us consider the brain as a machine with information-processing capabilities

- And look at different types of information-processing computations
  - **Unconscious processing (C0)**
  - **Global availability of information (C1)**: The selection of information for global broadcasting, making it flexibly available for computation and report
  - **Self-monitoring (C2)** of those computations, leading to a subjective sense of certainty or error

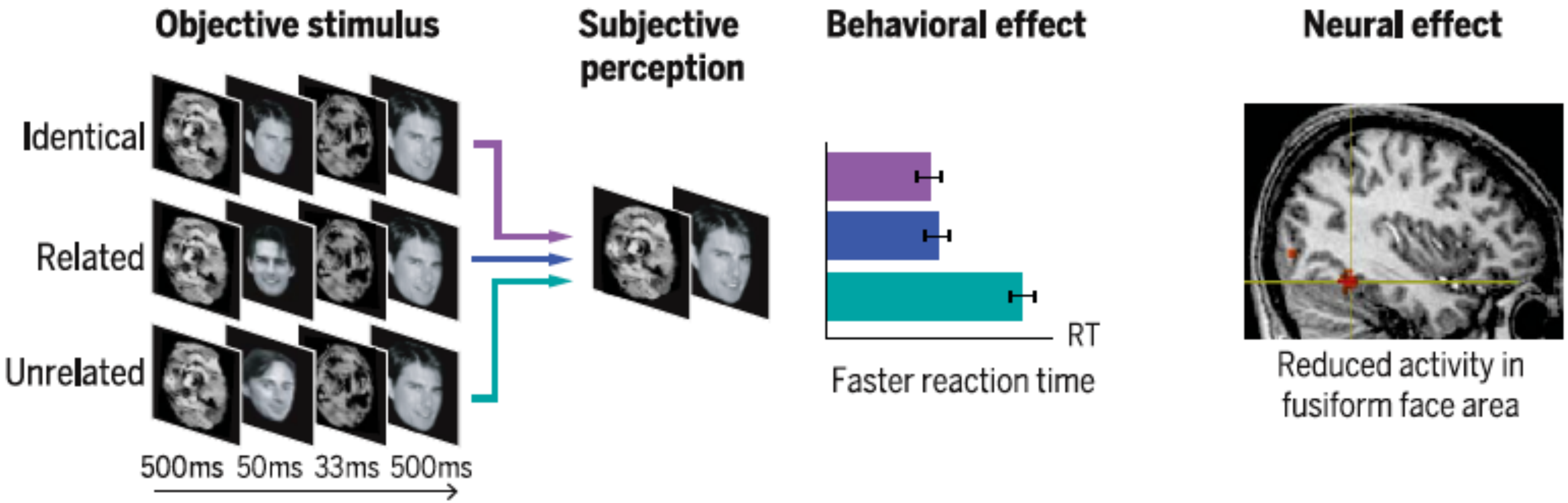
The multiple meanings of Consciousness

- Let us consider the brain as a machine with information-processing capabilities

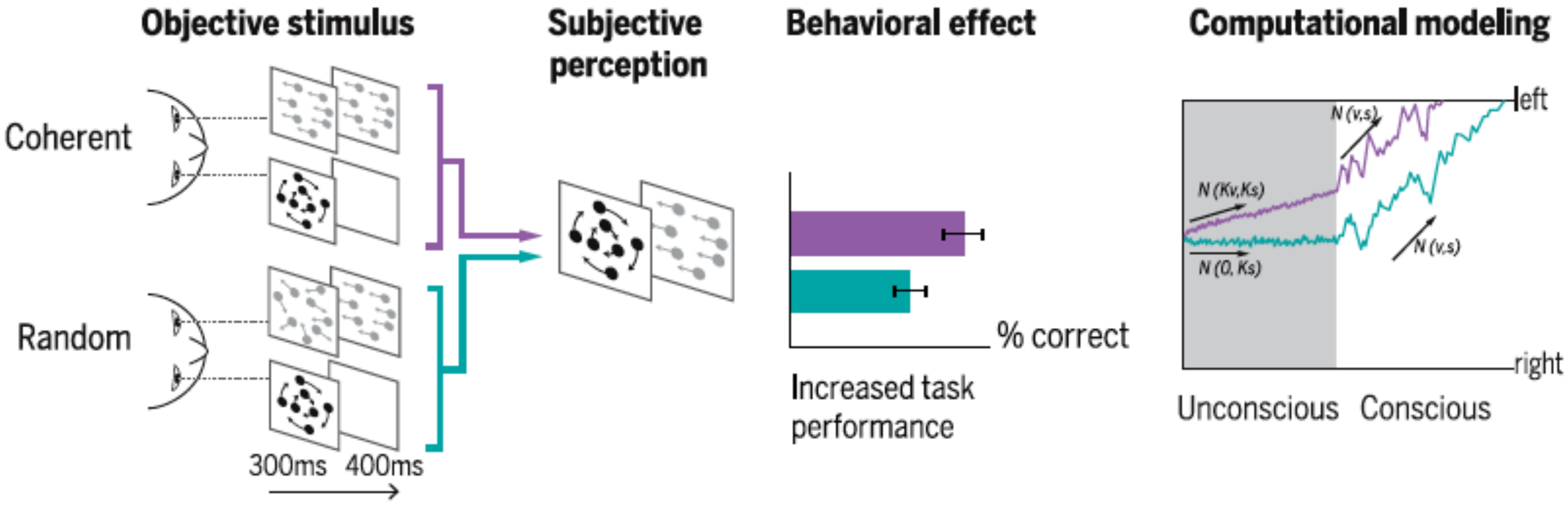
- And look at different types of information-processing computations
  - **Unconscious processing (C0)**
  - **Global availability of information (C1)**: The selection of information for global broadcasting, making it flexibly available for computation and report
  - **Self-monitoring (C2)** of those computations, leading to a subjective sense of certainty or error

- Paradigms to probe these types of computations:

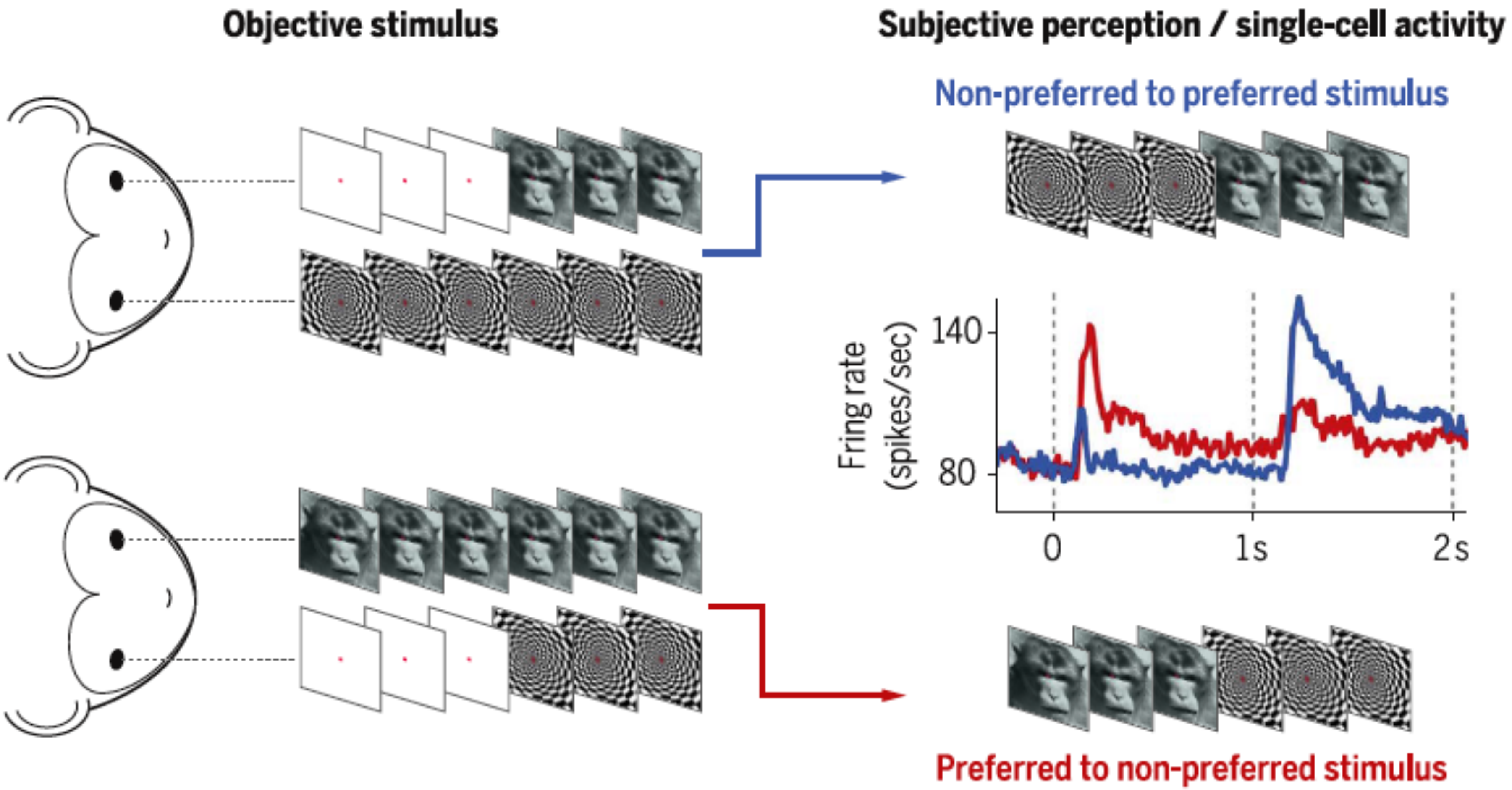
Unconscious processing (C0)



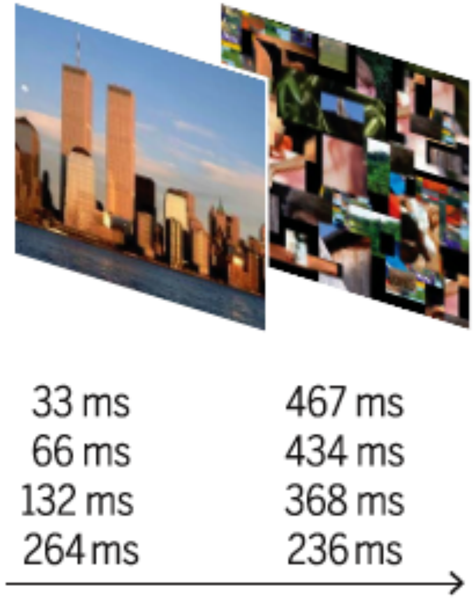
Unconscious processing (C0)

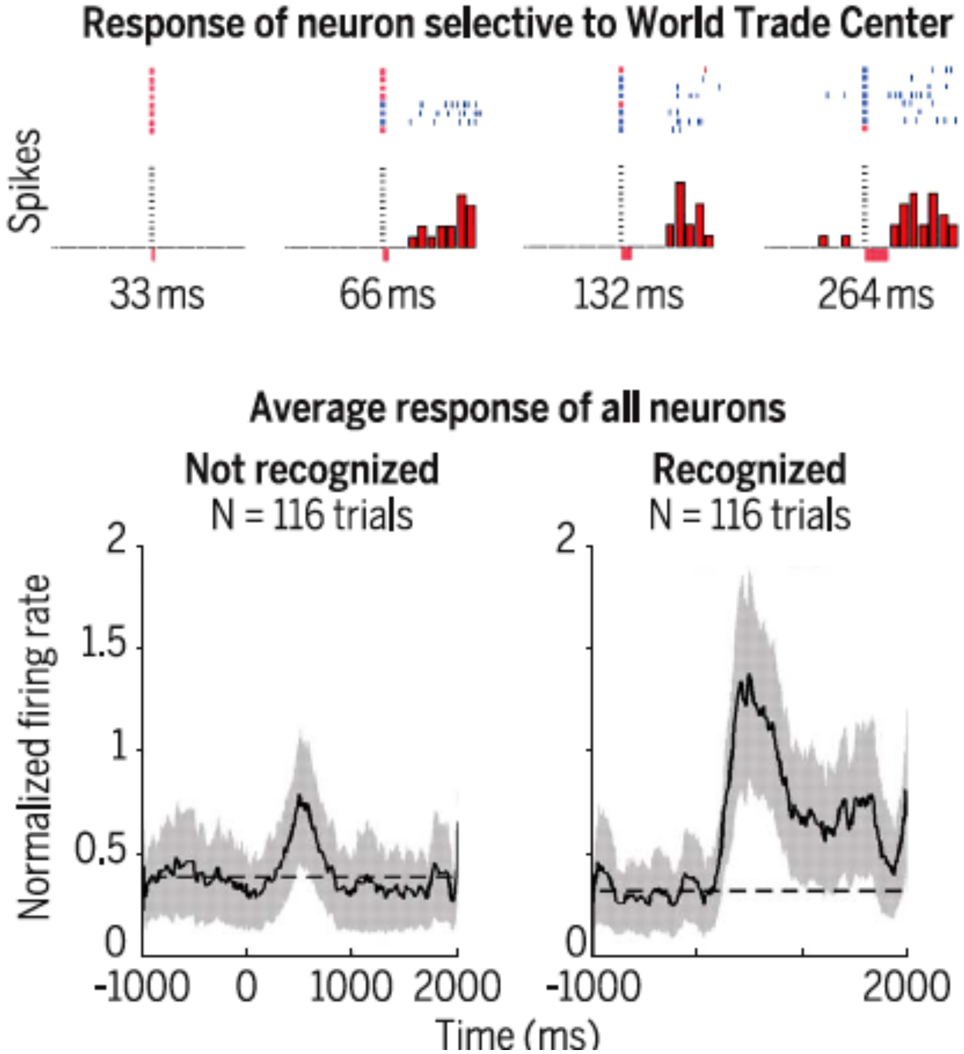


Global availability of information (C1)

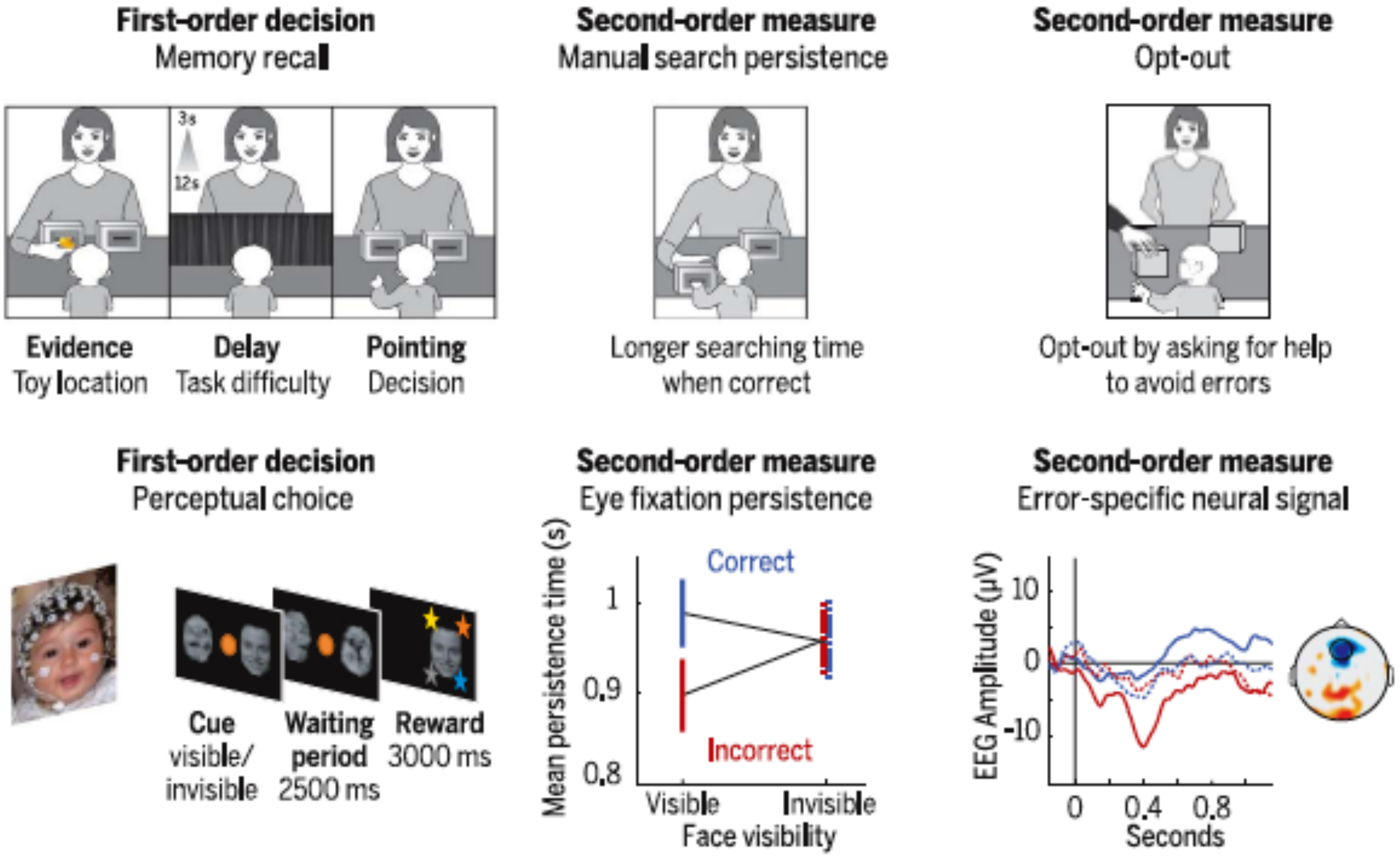


Global availability of information (C1)





Self-monitoring (C2)



Dissociation between C1 and C2

- They are largely orthogonal and complementary dimensions of what we call consciousness
- Self-monitoring can exist for unreportable stimuli (C2 without C1)
- Consciously reportable contents sometimes fail to be accompanied with an adequate sense of confidence (C1 without C2)

Synergy between C1 and C2

- Because C1 and C2 are orthogonal, their joint possession may have synergistic benefits to organisms
  - In one direction, bringing probabilistic metacognitive information (C2) into the global workspace (C1) allows it to be held over time, integrated into explicit long-term reflection, and shared with others
  - In the converse direction, the possession of an explicit repertoire of one's own abilities (C2) improves the efficiency with which C1 information is processed

Pathways to artificial consciousness

- What makes the difference to the processing related to C0 into non-conscious? What's needed to make it conscious?

- Is C1 sufficient?

- Is C2 sufficient?

- Is there a case of non-conscious processing with C1 AND C2?

- Is there any better alternative to C1 and C2 for AI?

Pathways to artificial consciousness

- Current machines are still mostly implementing computations that reflect unconscious processing (C0) in the human brain
- Endowing machines with global information availability (C1) would also allow the different modules to share information and collaborate to address impending problems
- To make optimal use of the information, it would also be useful for the machine to possess a database of its own states. Such self-monitoring (C2) would include an integrated image of itself as well as its internal databases

Pathways to artificial consciousness

- Combining C1 and C2 in adversarial learning

Pathways to artificial consciousness

- Adversarial learning, involves having a secondary network “compete” against a generative network so as to critically evaluate the authenticity of self-generated representations

- When reality monitoring (C2) is coupled with C1 mechanisms, the resulting machine may more closely mimic human consciousness in terms of affording global access to perceptual representations while having an immediate sense that their content is a genuine

- How to do it in practice?

Pathways to artificial consciousness

- Using a generative adversarial network (GAN): One network generates candidates (generative) and the other evaluates them (discriminative)

Pathways to artificial consciousness

- Using a generative adversarial network (GAN): One network generates candidates (generative) and the other evaluates them (discriminative)

    - What is a discriminative model?
    - What is a generative model?

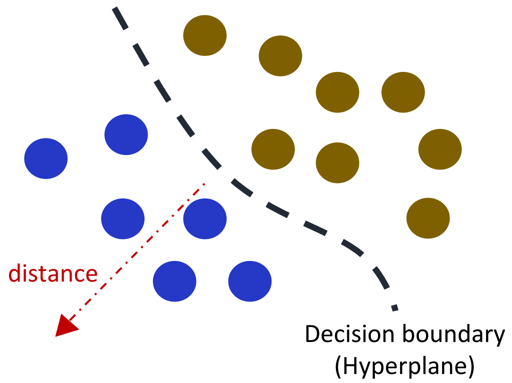
Pathways to artificial consciousness

- Using a generative adversarial network (GAN): One network generates candidates (generative) and the other evaluates them (discriminative)

Discriminative models learn the boundary between classes

Examples:

- Logistic regression
- SVM
- NN



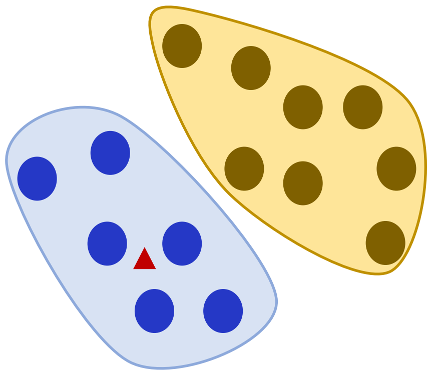
Pathways to artificial consciousness

- Using a generative adversarial network (GAN): One network generates candidates (generative) and the other evaluates them (discriminative)

Generative models model the distribution of individual classes

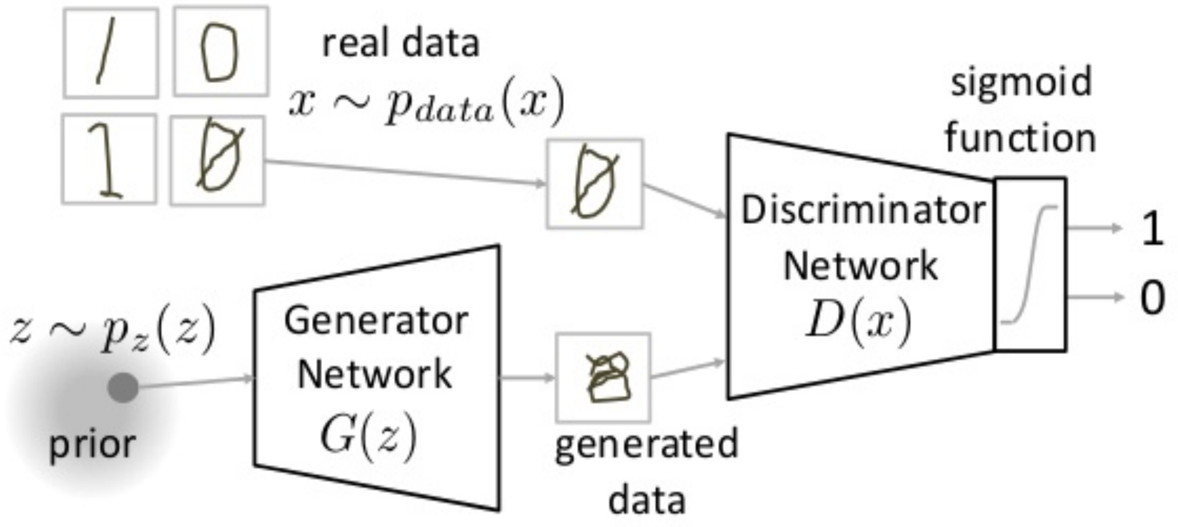
Examples:

- Naïve Bayes
- Gaussian Discriminant Analysis



Pathways to artificial consciousness

- Using a generative adversarial network (GAN): One network generates candidates (generative) and the other evaluates them (discriminative)
  - Example

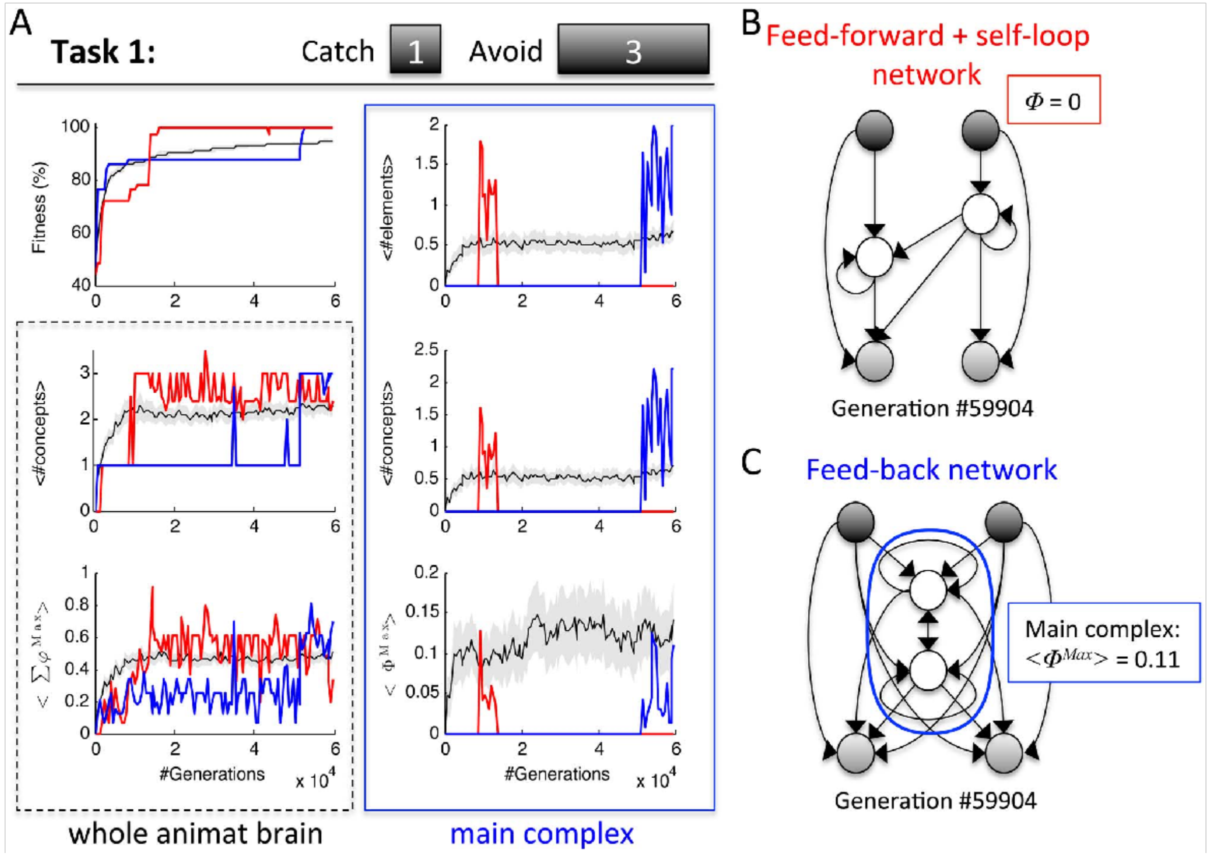


Pathways to artificial consciousness

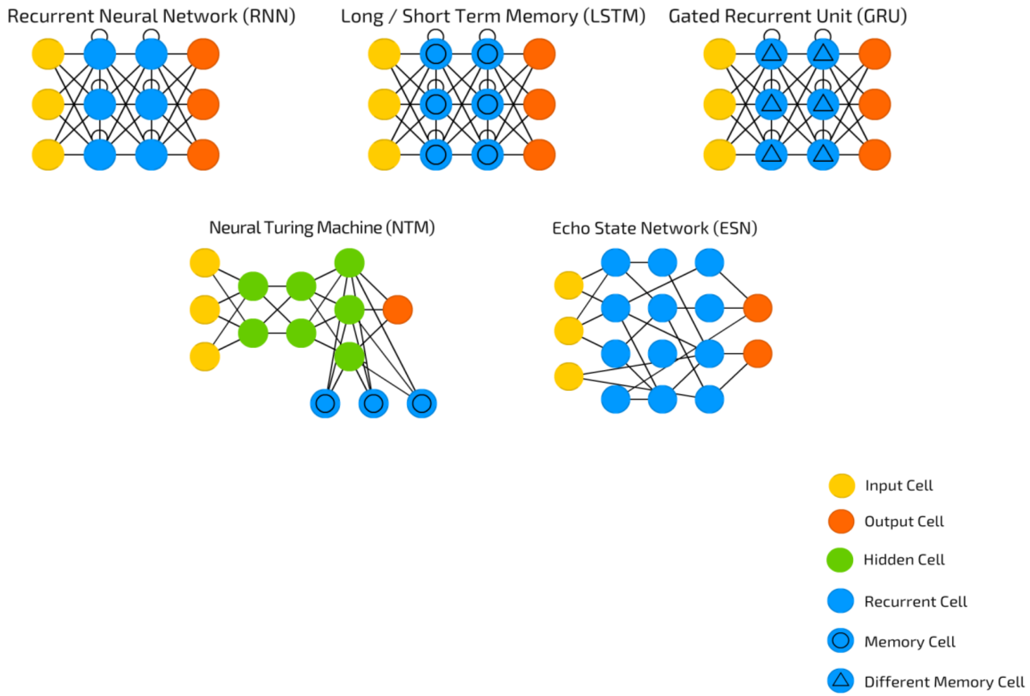
- Other ways?
    2. Optimizing integration in animats through evolution (IIT)
    3. Minimizing error (Predictive coding)

Pathways to artificial consciousness

- Optimizing integration in animats through evolution (IIT)



Feedback networks are the key?

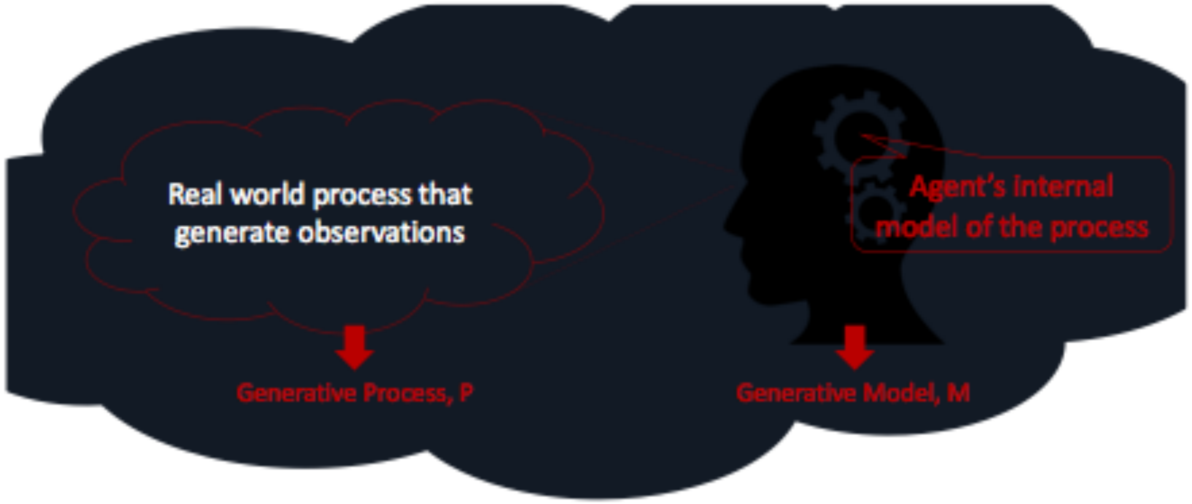


Pathways to artificial consciousness

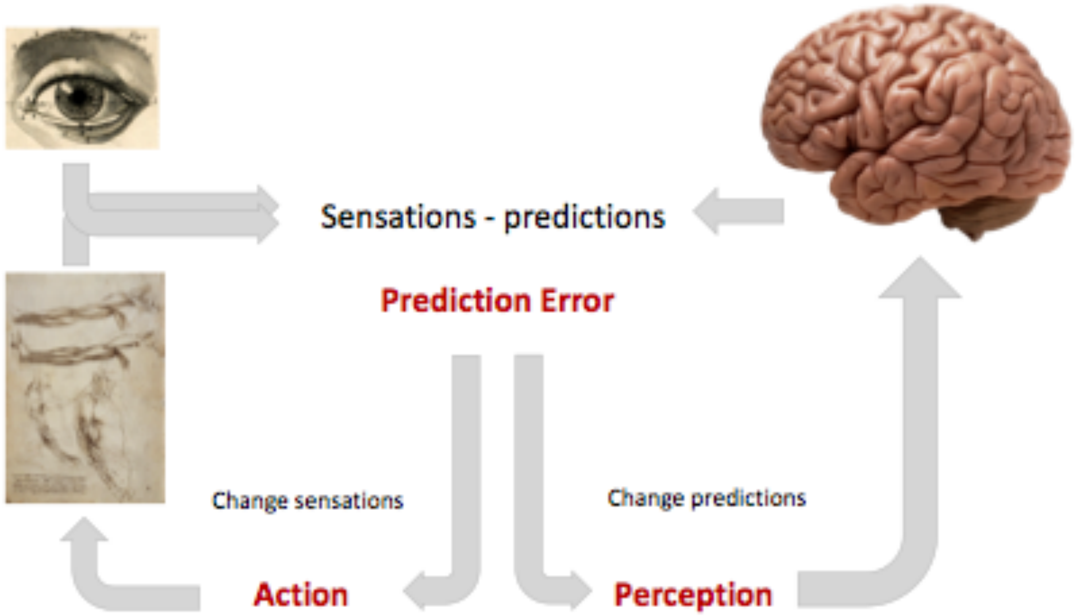
- Minimizing error between two generative models
  - Error detection provides a particularly clear example of self-monitoring; just after responding, we sometimes realize that we made an error and change our mind

Pathways to artificial consciousness

- Minimizing error between two generative models



Sensory observations generated by P are observed by the agent while the agent is acting on the world to change P



Summary

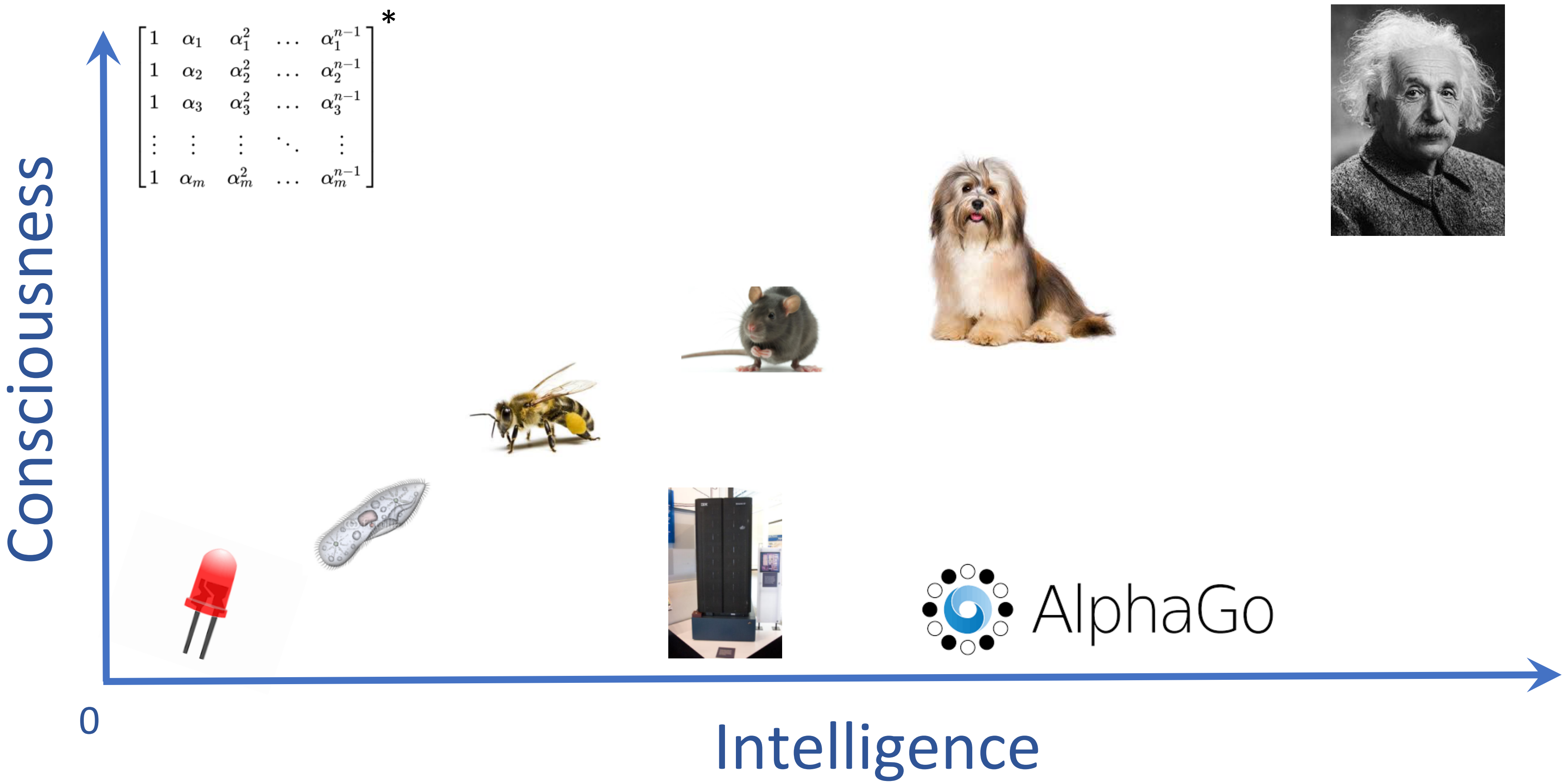
- The human brain as blueprint for artificial consciousness
- The multiple meanings of consciousness: C0, C1, C2
- Generative models and the ability to reflexively represent oneself

Discussion

- Intelligence = consciousness?

Discussion

- Intelligence vs. Consciousness



$$\begin{bmatrix} 1 & \alpha_1 & \alpha_1^2 & \dots & \alpha_1^{n-1} \\ 1 & \alpha_2 & \alpha_2^2 & \dots & \alpha_2^{n-1} \\ 1 & \alpha_3 & \alpha_3^2 & \dots & \alpha_3^{n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \alpha_m & \alpha_m^2 & \dots & \alpha_m^{n-1} \end{bmatrix}^{*}$$

(*) According to IIT

Discussion

- Does one give rise to the other?

- Measuring intelligence and consciousness
  - IQ (humans), fitness (animat example), utility (artificial agents), etc.
  - Phi, BoC, etc.

- Access/Phenomenal consciousness and C1/C2

References

1. Dehaene, Stanislas, Hakwan Lau, and Sid Kouider. "What is consciousness, and could machines have it?." Science (2017): 486-492.
2. Kim, Chai-Youn, and Randolph Blake. "Psychophysical magic: rendering the visible ‘invisible’." Trends in cognitive sciences 9.8 (2005): 381-388.
3. LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." nature 521.7553 (2015): 436.
4. 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 Computational Biology, 10*(12).
5. Friston, Karl, et al. "Active inference: a process theory." Neural Computation 29.1 (2017): 1-49.