ARAYA

Consciousness & Intelligence

Ryota Kanai (Araya, Inc.)

ARAYA

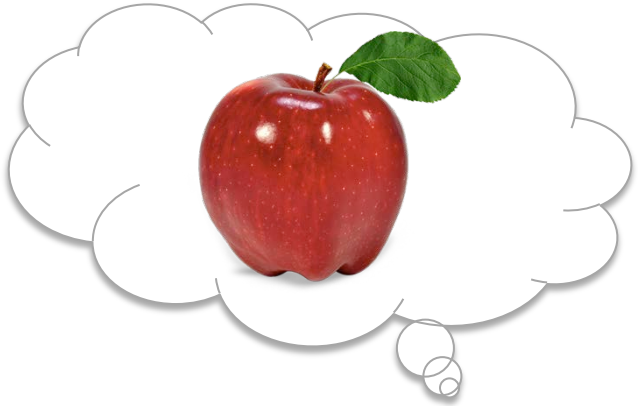
Introduction

5 Problems of Consciousness

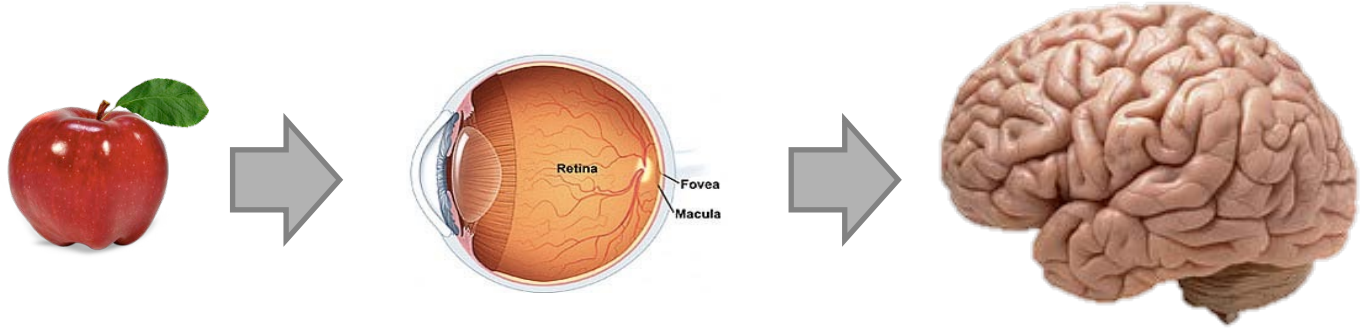
Problem 1: The Hard Problem

Why does subjective experience occur at all?

Subjective Experience

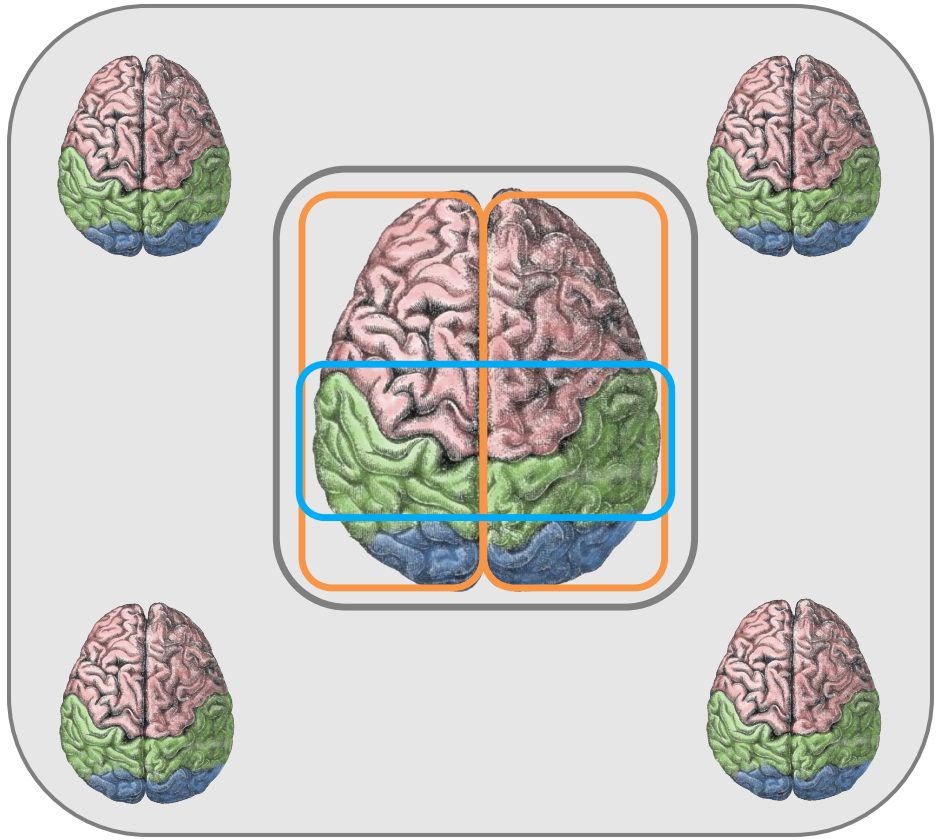


Physical World



Problem 2: The Boundary Problem

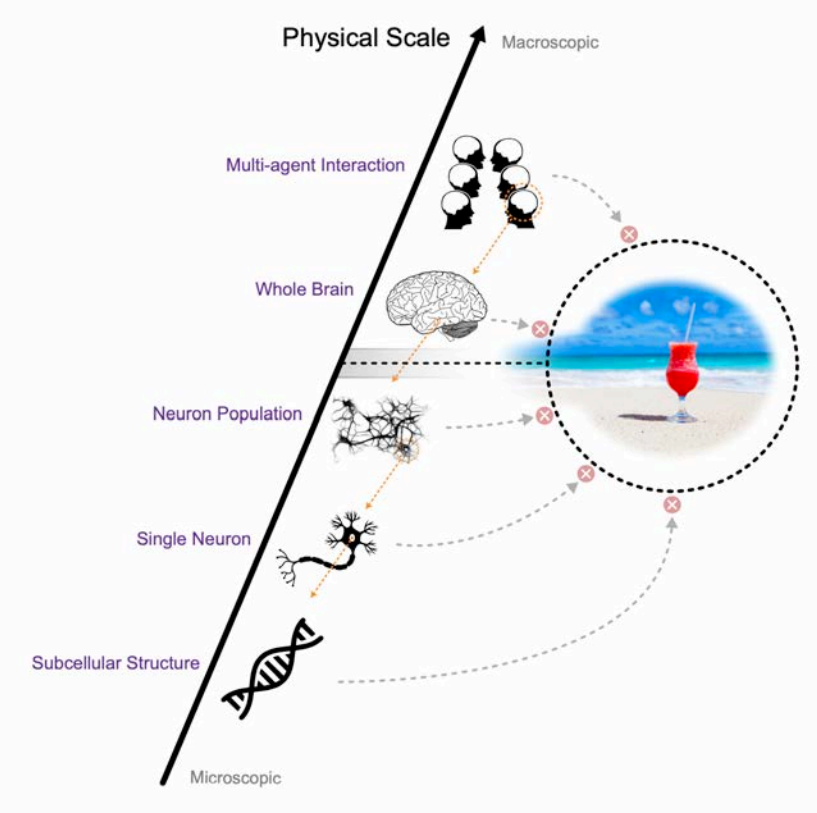
What determines the boundary of consciousness?



- Brains have separate consciousness
- Not all parts of the brain are included in consciousness

Problem 3: The Scale Problem

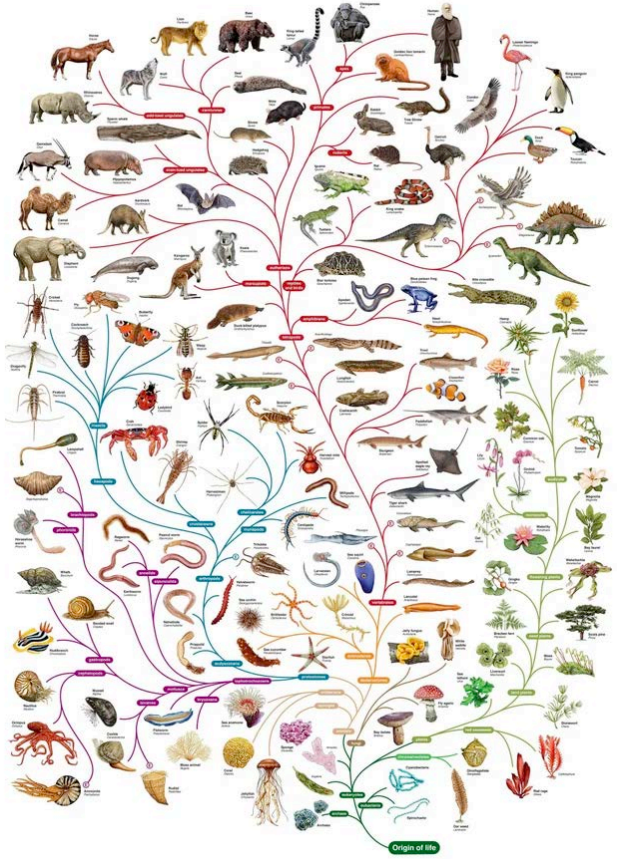
At what scale does consciousness occur?



Contents of consciousness appear to correspond to information coded at a particular scale of neuronal activities.

Problem 4: The Function Problem

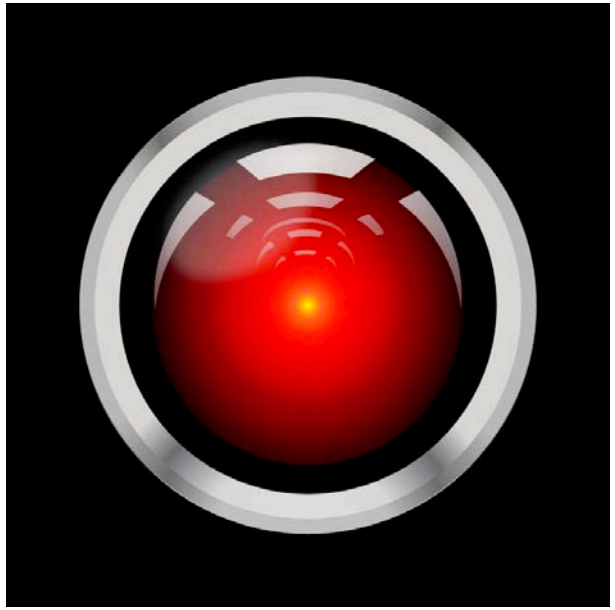
What is the function of consciousness?

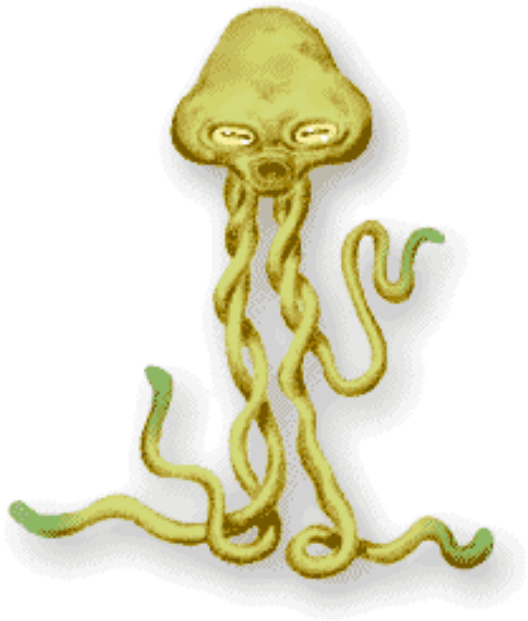


Has consciousness evolved for a reason?

Problem 5: The Existence Problem

Can we prove consciousness?





How can we determine whether machines or aliens have conscious experience?

ARAYA

Consciousness and Artificial General Intelligence

(Warning: lots of speculations)

Motivation: Consciousness and Intelligence

Question

In science fiction, AI wakens to the self-awareness and exhibits transcendence to a different level of intelligence. This wakening of AI presupposes a form of criticality of intelligence where AI gains consciousness as a consequence of improvements in function.

→ Is consciousness and intelligence independent? Or do they have any common root?

Epiphenomenalism

With our current scientific understanding of the reality, there is no causal role for conscious experience (no room for substance dualism). Scientists seem to take the view of epiphenomenalism (implicitly or explicitly).

Confusion about two questions about functions of consciousness

1. Does the subject experience (qualia) play any function?
   - This is about the causal role of qualia.
2. What does consciousness enable an agent to do?
   - This only concerns the functional consequence of some information giving rise to consciousness.
   - Also called the Hard Question (c.f. Dennett)

Main Claim about Consciousness and Intelligence

Claim

- Consciousness works as a platform for general intelligence.

Reason

- Putative functions of consciousness serve the purpose of general intelligence.

Defining General Intelligence

Definition: Generality of Intelligence

Generality of intelligence is measured as the ability to efficiently solve multiple tasks, including tasks novel to the agent, using knowledge and models learned through past experiences.

Note

- With this definition, generality of intelligence is essentially the efficiency of transfer learning/meta-learning.
- Some other qualities may be important for intelligence (e.g. symbol manipulations etc.)
- Sometimes AGI is defined in terms of human-level intelligence, but it's too vague to be useful (we can't quantify).

c.f. Chollet (2019)

The intelligence of a system is a measure of its skill-acquisition efficiency over a scope of tasks, with respect to priors, experience, and generalization difficulty.

Three approaches to building AGI

We propose three ways to approach construction of AGI.

Solution by simulation

- If you've acquired a forward model (world model), we can use internal simulations to adaptively find a solution to reach new goals without further sampling from the environment. In this method, even if the objectives change, the strategy to reach them are derived on the spot.
- The models learned in the past are used to solve a wide variety of future problems.

Solution by combination

- We have several task-specific models that we have learned in the past, and we can flexibly combine them to solve new and diverse tasks efficiently.

Solution by generation

- We can create a latent space for embedding (representing) neural networks, and from that space, a new network can be generated to solve new tasks.

Three approaches to building AGI (1)

Solution by simulation

- If a forward model (world model) has been successfully learned, we can use the model to run internal simulations to adaptively find a solution to reach new goals without further sampling from the environment.
- In this method, even if the objectives change, the strategy to reach them are derived on the spot. The models learned in the past are used to solve a wide variety of future problems.

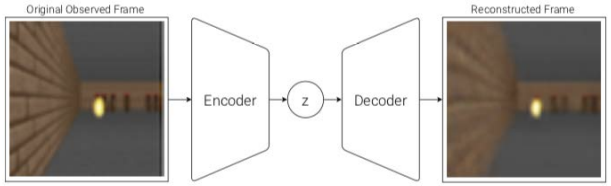
Corresponding Theory of Consciousness

Information Generation Hypothesis (Kanai et al., 2019)

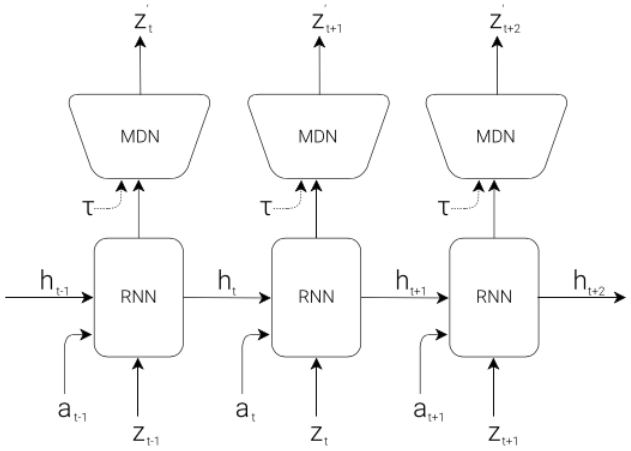
The function of consciousness is to generate information using the models constructed via interactions with the environment. Specifically, this allows agents to interact with counterfactual situations detached from the current sensory input. Using models in this way, it becomes possible to perform action planning based on simulation of the future and to imagine situations that do not actually occur.

World Models (Ha & Schmidhuber, 2018)

Vision: VAE



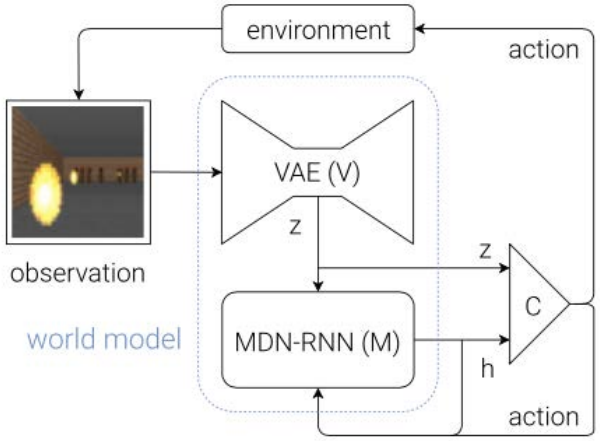
Memory: MDN-RNN



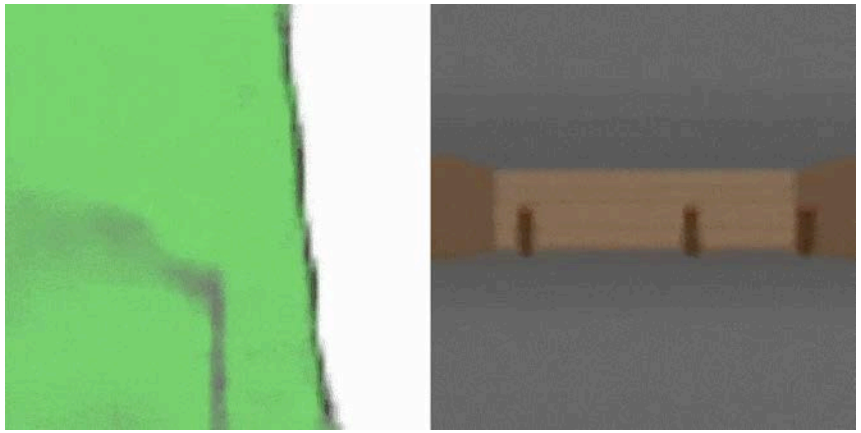
Action: Very simple linear model

$$a_t = W_c \left[z_t \; h_t \right] + b_c$$

Everything together



Mental simulation (training in hallucination)

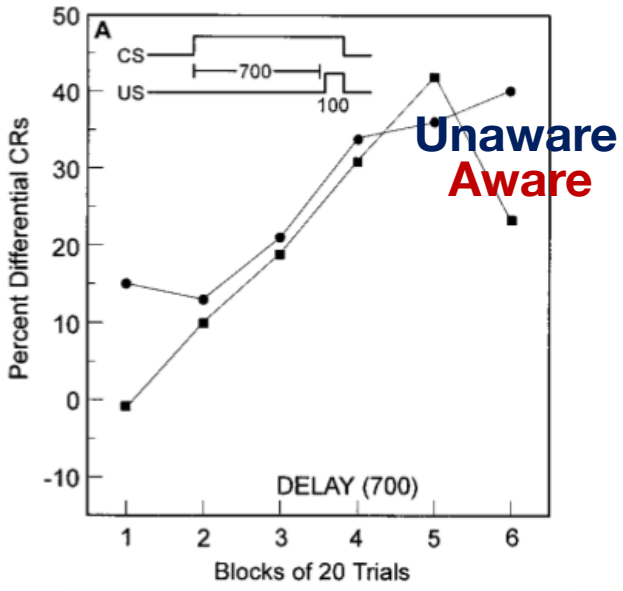


Trace conditioning

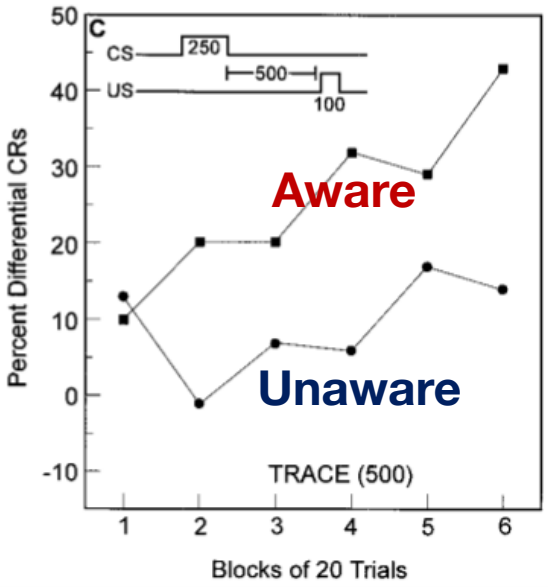
Delay conditioning occurs without awareness

Trace conditioning requires awareness (and hippocampus)

Delay Conditioning



Trace Conditioning

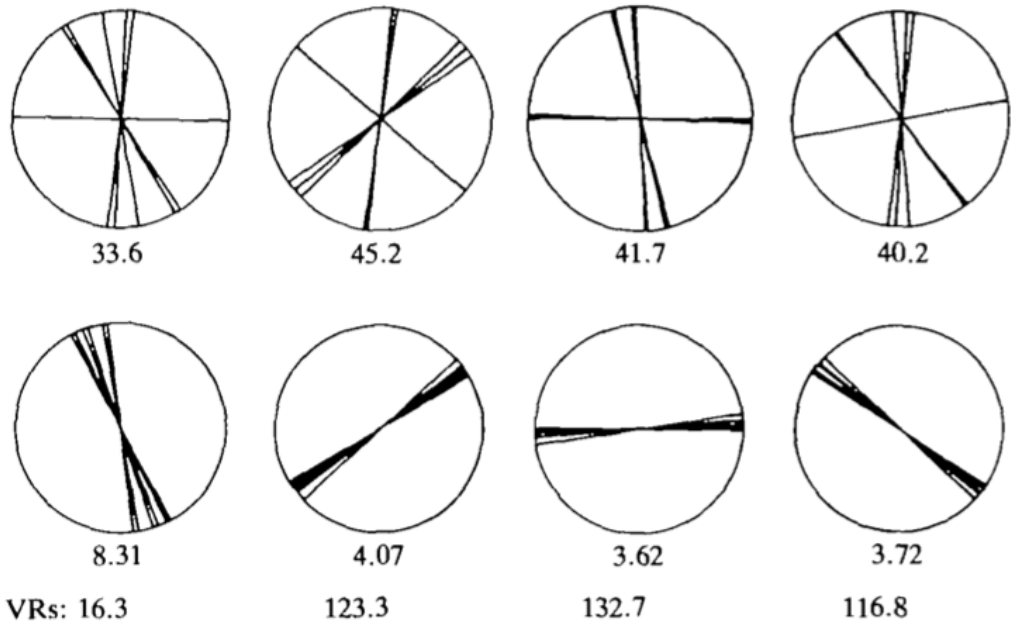


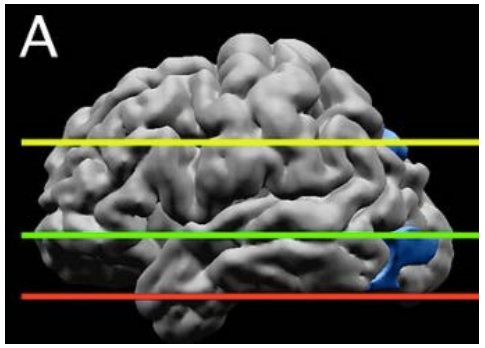
Clark & Squire (1998)

Online and memory guided action

Perceptual judgement

Visually guided action





Memory guided action depends on conscious perception.

Milner et al. (1991)

Information Generation Hypothesis

Function of Consciousness = The ability to generate representations of events disconnected from the present environment.

A model of the world, and the agent's sensory motor contingency is required for the ability to simulate the world internally.

Intention/Planning/Imagination (i.e. mental simulation)

- Counterfactual predictions of the consequences of future actions.

Non-Reflexive Behaviour

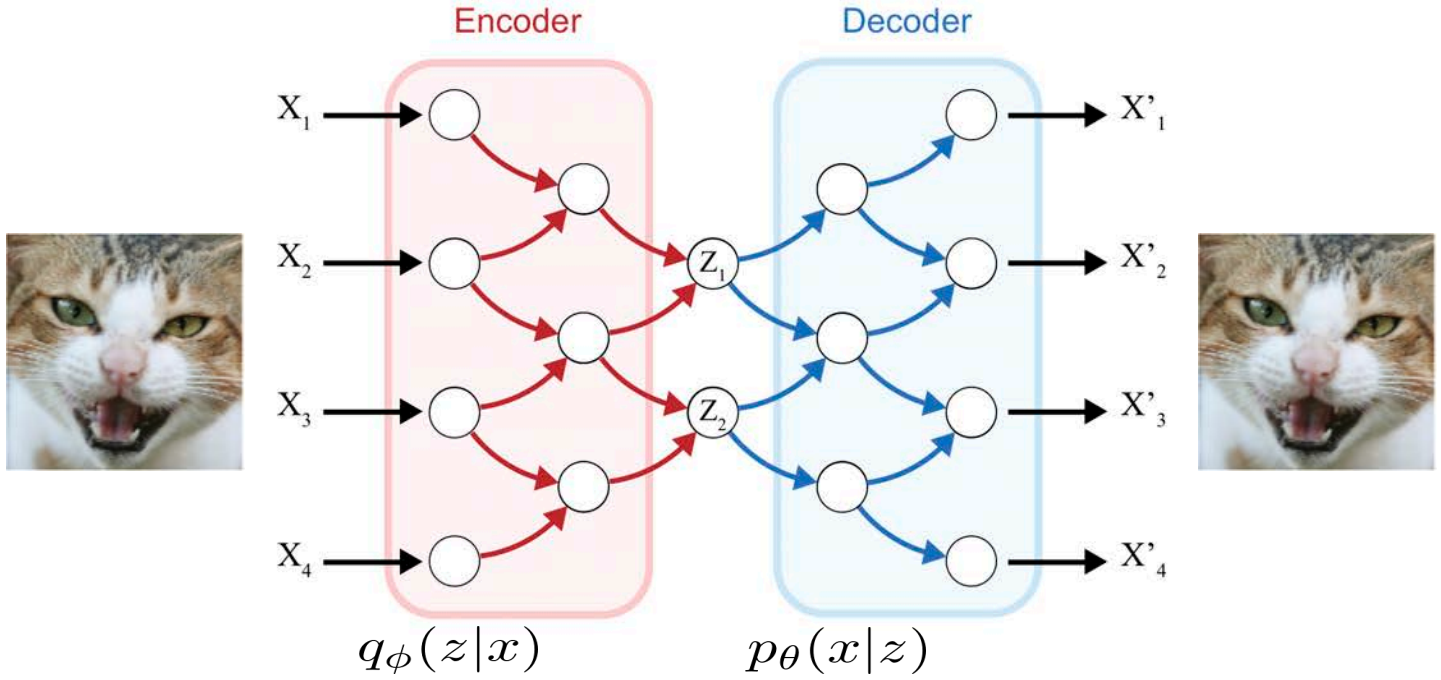
- Mental simulation detached from the present external stimuli.

Short-term memory

- Interaction with sensory information from the past.

What does it mean to generate information?

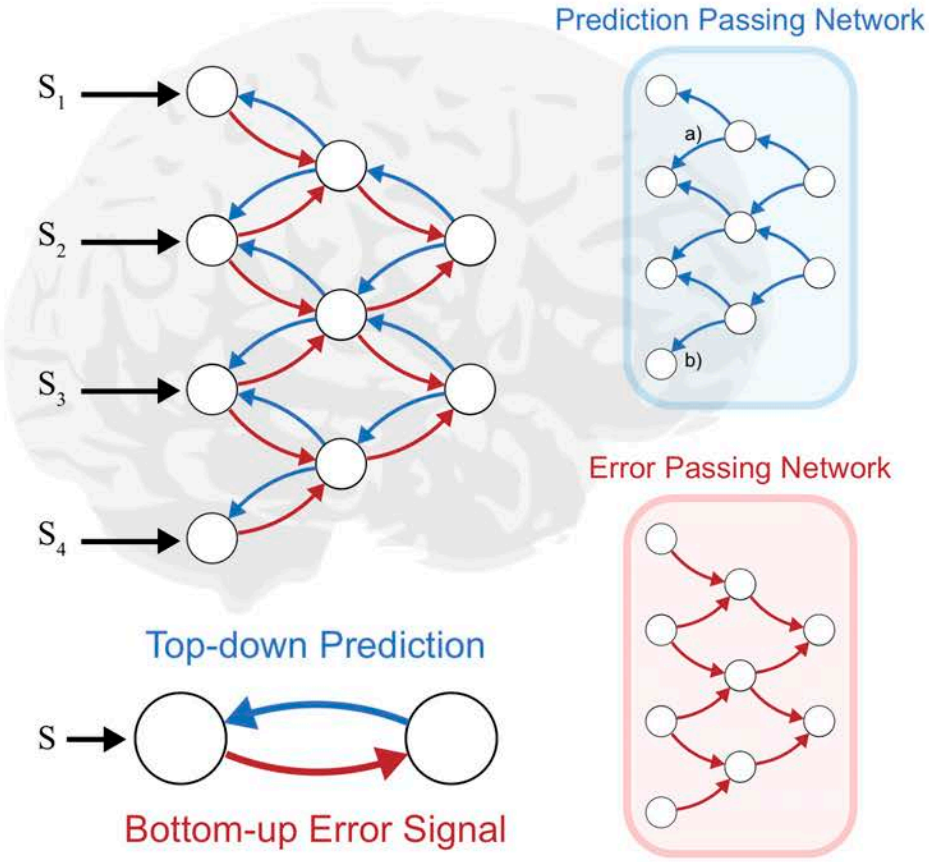
a) Autoencoder



Info Processing — Unconscious

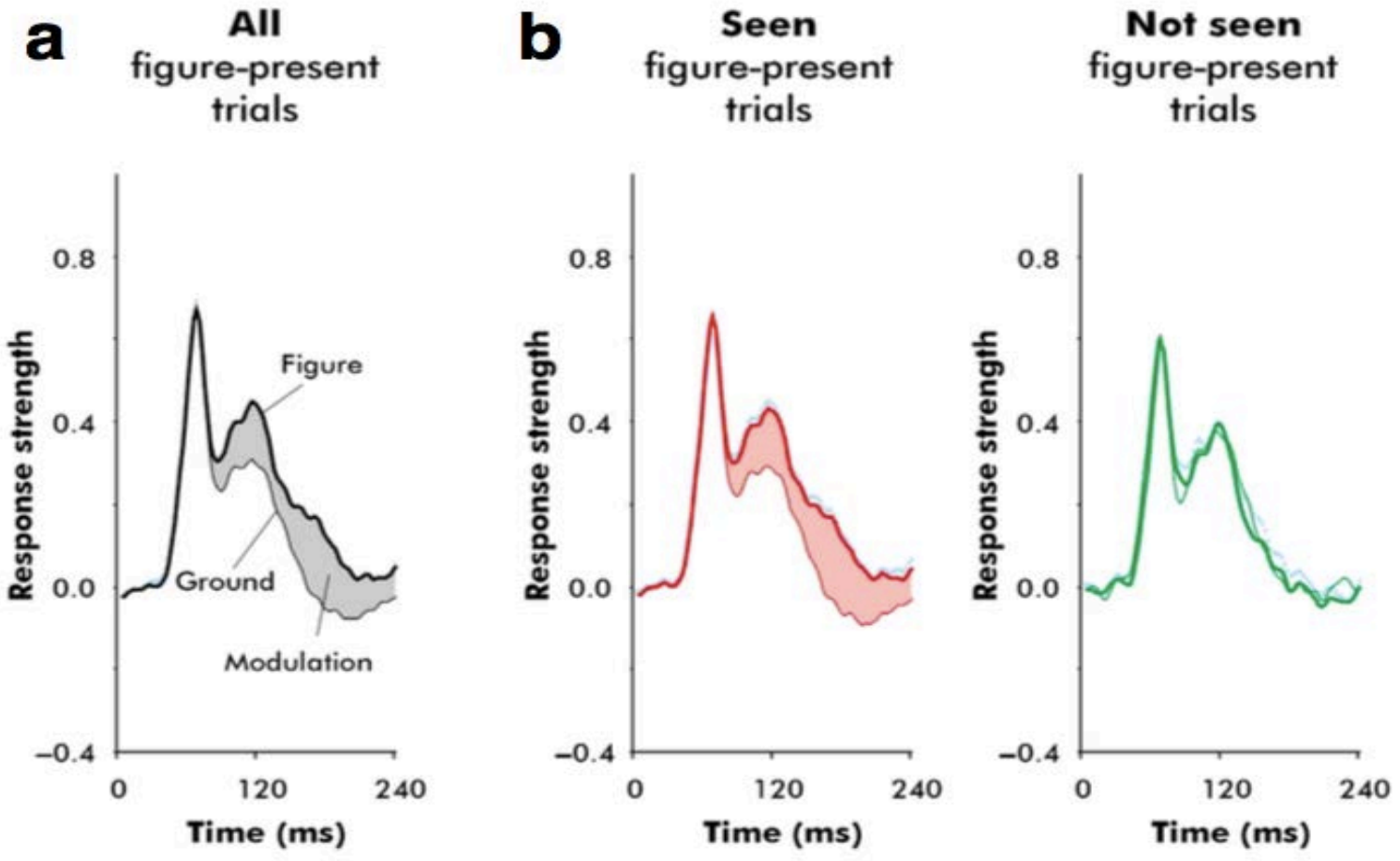
Info Generation — Conscious

b) Predictive Coding



Loss function: $\log P(X) - D_{KL}[Q(z|X)\|P(z|X)] = E[\log P(X|z)] - D_{KL}[Q(z|X)\|P(z)]$ *(Free energy)*

Feedback and visual awareness



Delayed activity is observed when the animal reported visibility.

Super et al. (2001)

Three approaches to building AGI (2)

ARAYA

Solution by Combination

- We have several task-specific models that we have learned in the past, and we can flexibly combine them to solve new and diverse tasks efficiently.

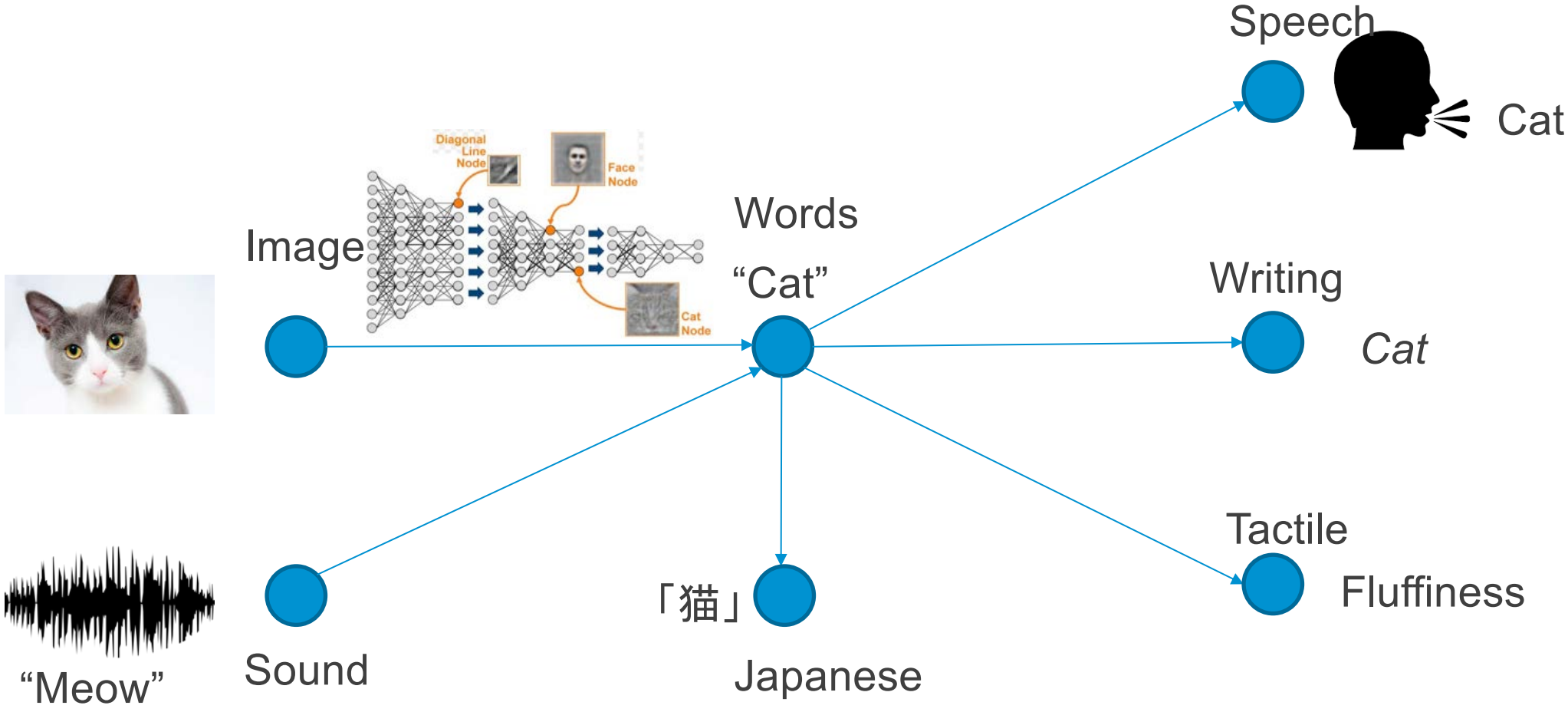
Corresponding Theory of Consciousness

Global Workspace Theory (Baars, 1988, 1997, 2002; Dehaene et al., 2003) as a Shared Latent Space.

The function of consciousness is to provide compatibility of data across models by connecting the latent spaces of many function-specific models. Once compatibility in the latent space is established, it is possible to create new functions instantly by flexibly combining multiple models. The difference between the part that contributes to the content of consciousness and the part that does not contribute to the content of consciousness in the brain, that is, the inside and outside of the global workspace is determined by whether or not it is included in the scope of this shared latent space. This is a re-interpretation and extension of the original global workspace.

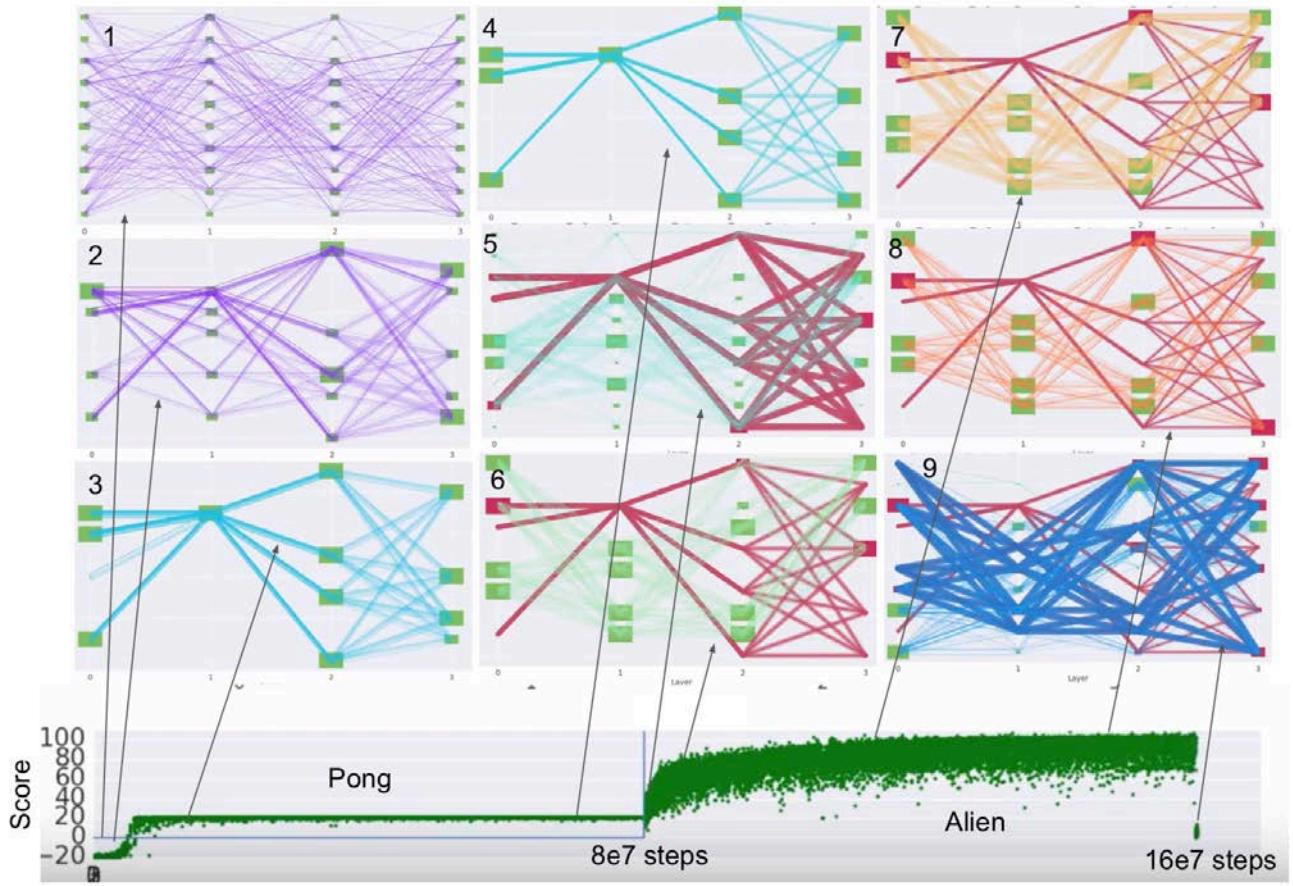
Network of Networks

New problems can be solved by finding a path in a network of networks.

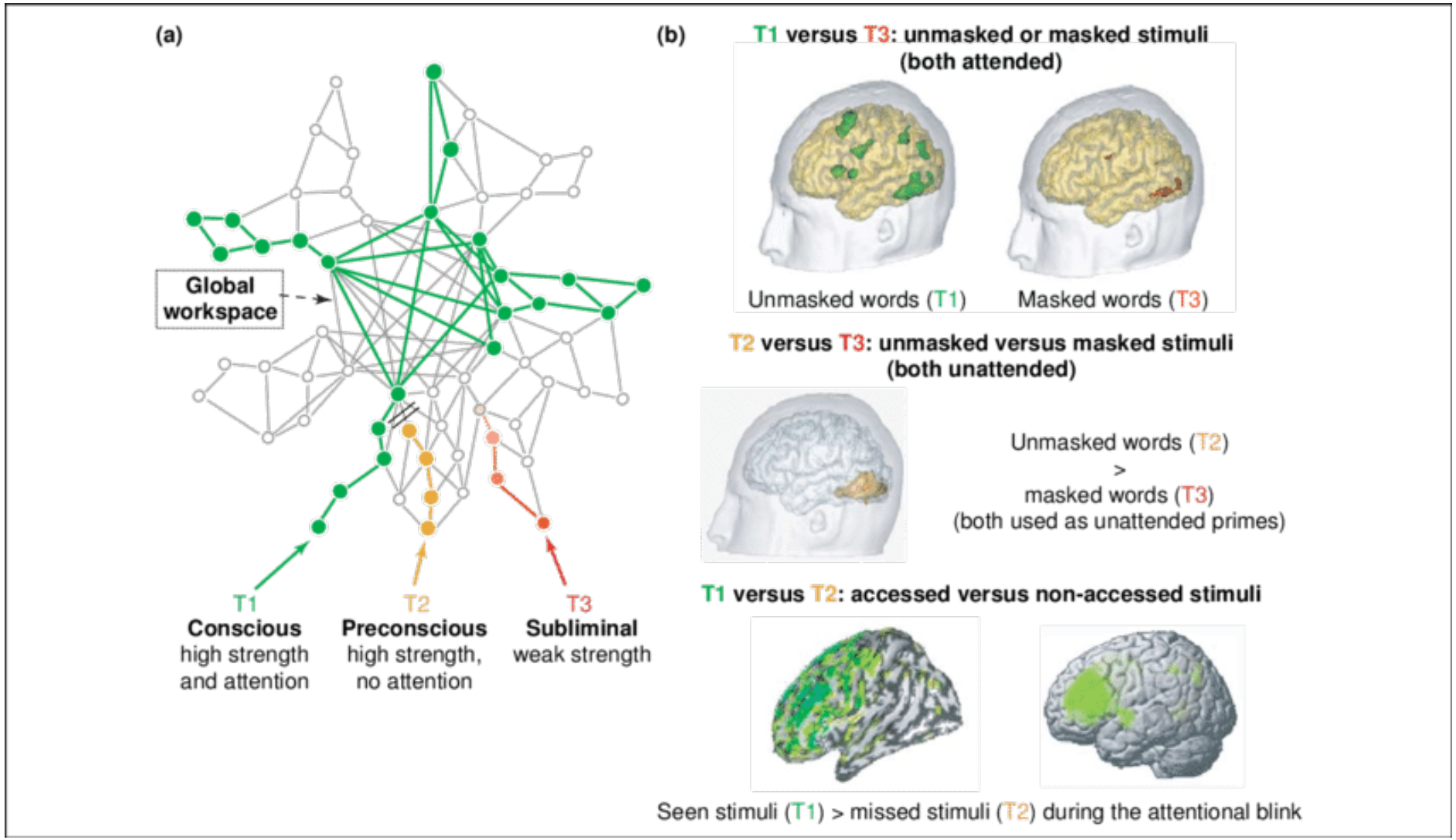


Solution by Combination: PathNet

PathNet is conceptually relevant for the idea of solution by combination as it solves problems by combining pretrained networks.

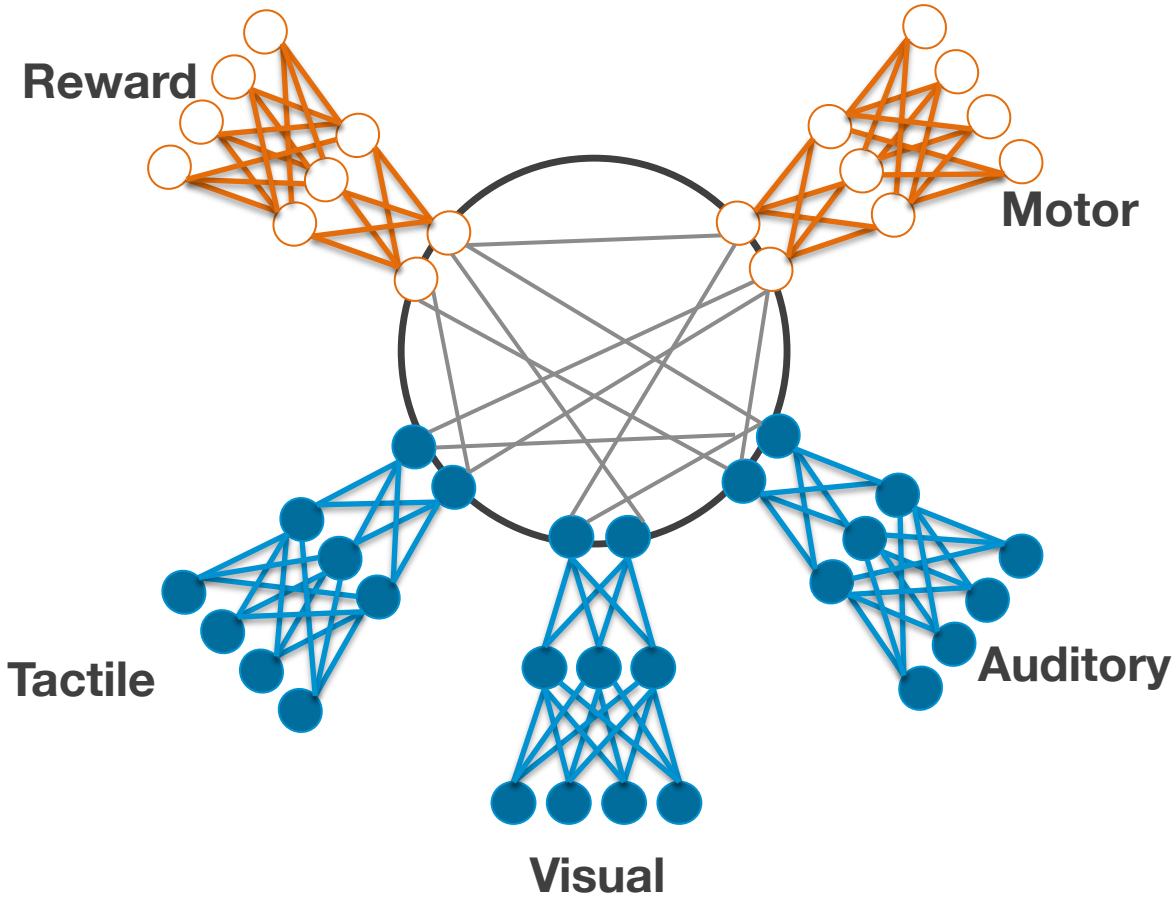


Global Workspace Theory



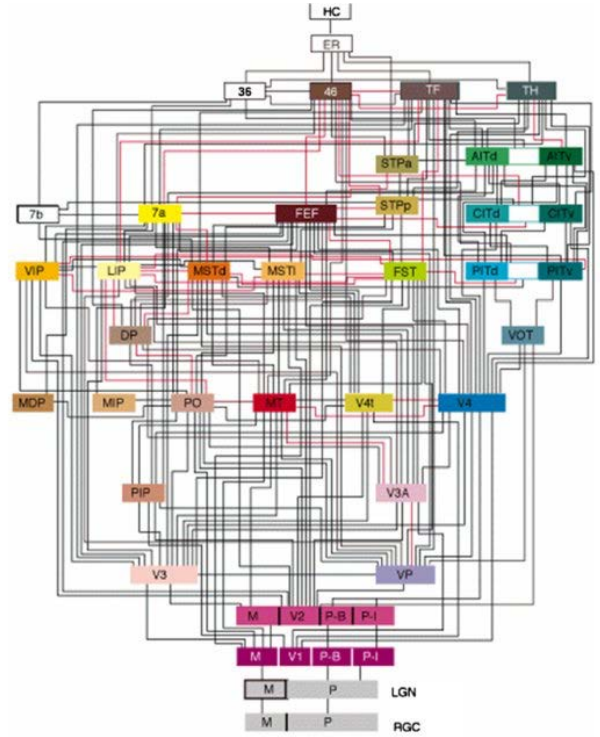
Global Workspace as a Shared Latent Space

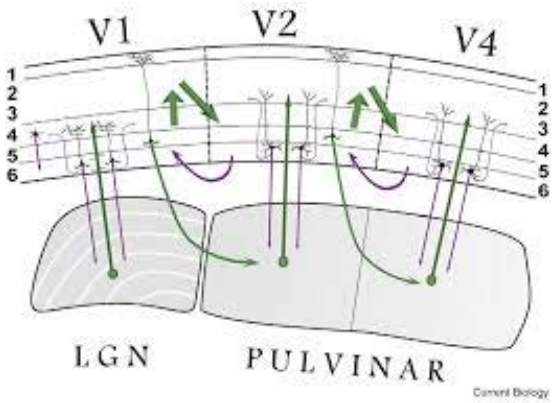
- To achieve solution by combination, the input and output formats must be compatible across models to make use of different types of data.
- In the global workspace theory of consciousness, consciousness enables a flexible combination of specialized modules, and as such should be regarded as as a platform that allows combinations of multiple models.

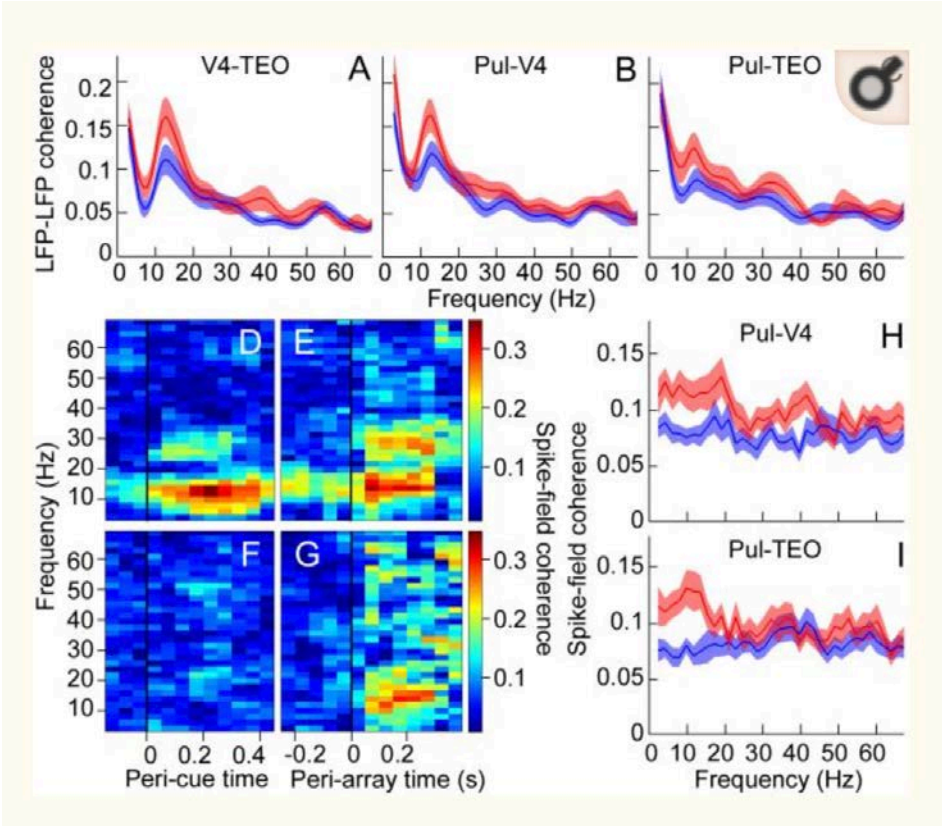


Brain implementation of routing in a netrok

A possible neural substrate for learning input-output relationships between cortical regions is thalamo-cortico-thalamic circuits.







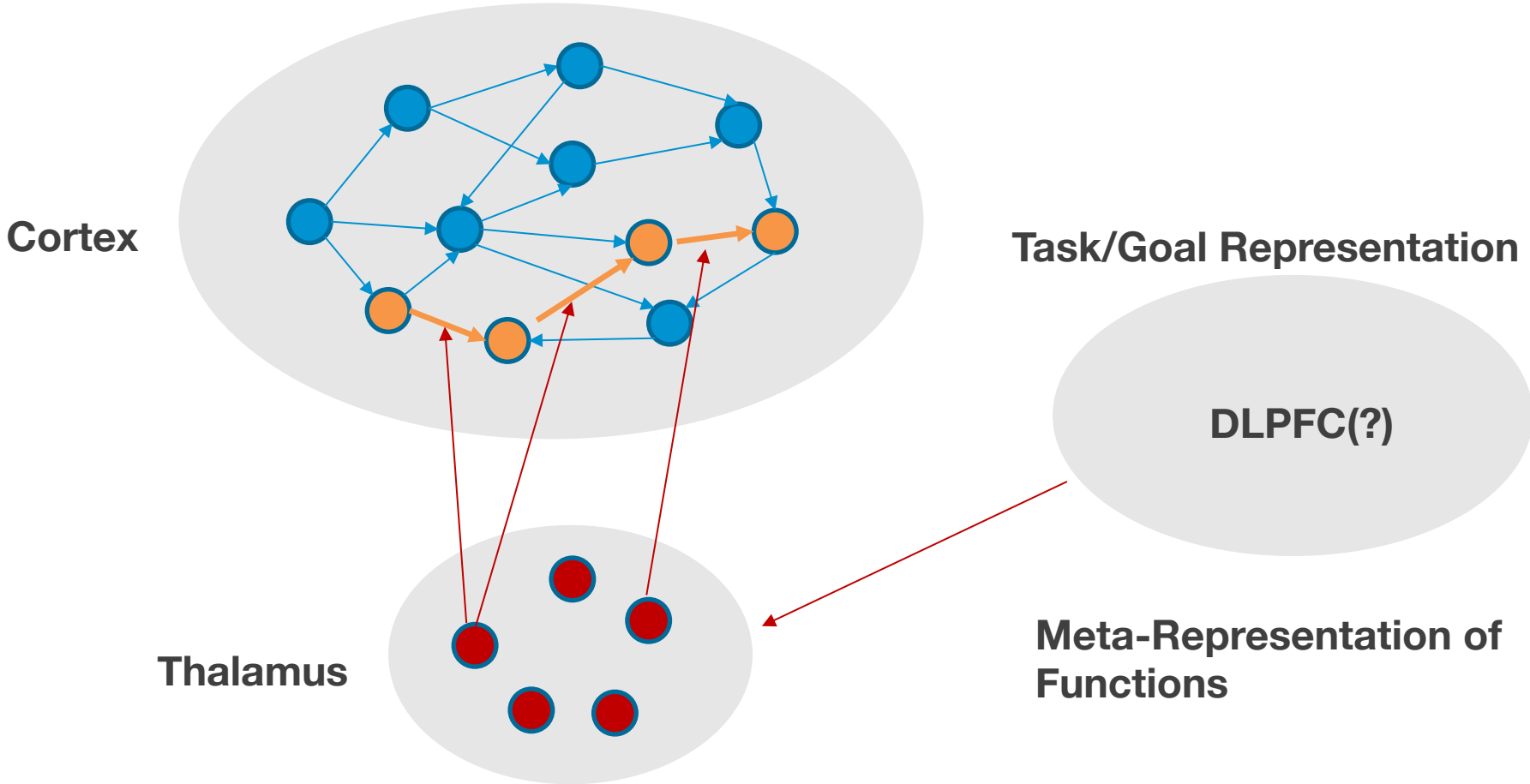
Each connection corresponds to a network

Pulvinar seems to dynamically gate between cortical regions.

Brain implementation of routing in a network

A schematics of how the thalamus creates new networks on the fly.

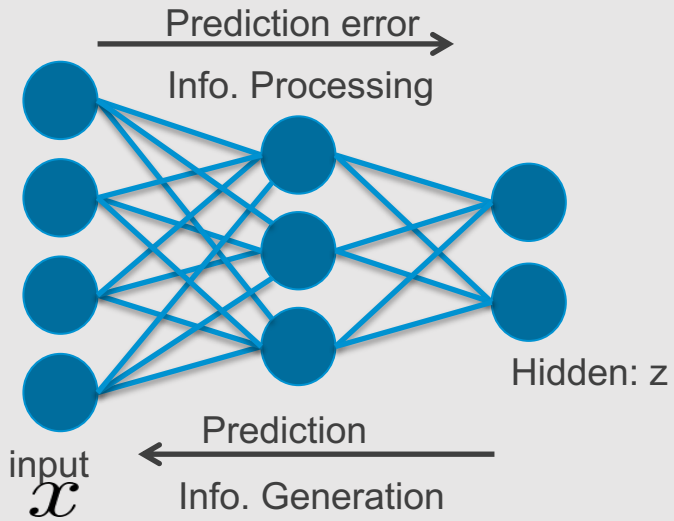
This is how the brain might achieve a broad range of functions via combination.



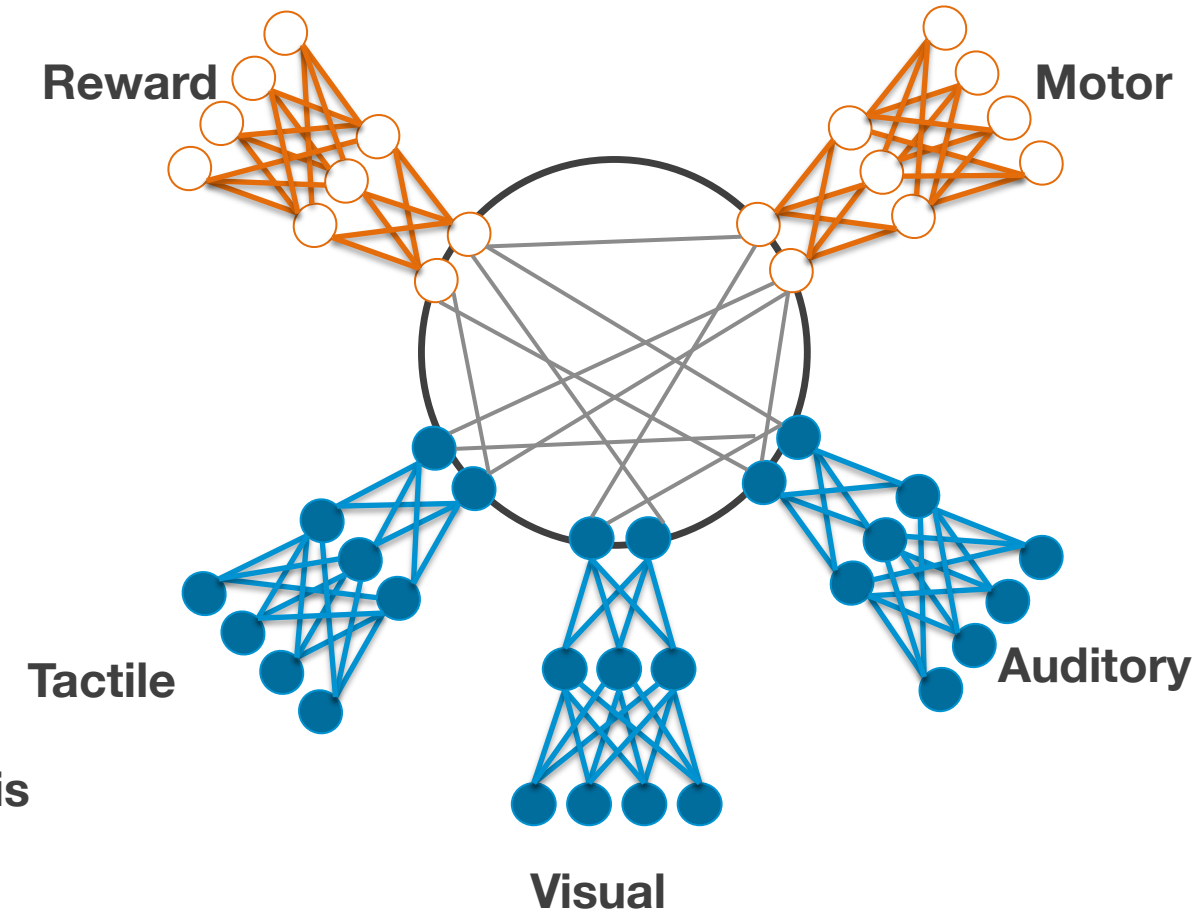
A global workspace as a latent space platform

ARAYA

A Predictive Coding Unit



Connecting latent space to establish a common language



Consciousness might be the latent space that is connected across modalities via a common language.

Three approaches to building AGI (2)

ARAYA

Solution by generation

- We can create a latent space for embedding (representing) neural networks, and from that space, a new network can be generated to solve new tasks.

Corresponding Theory of Consciousness

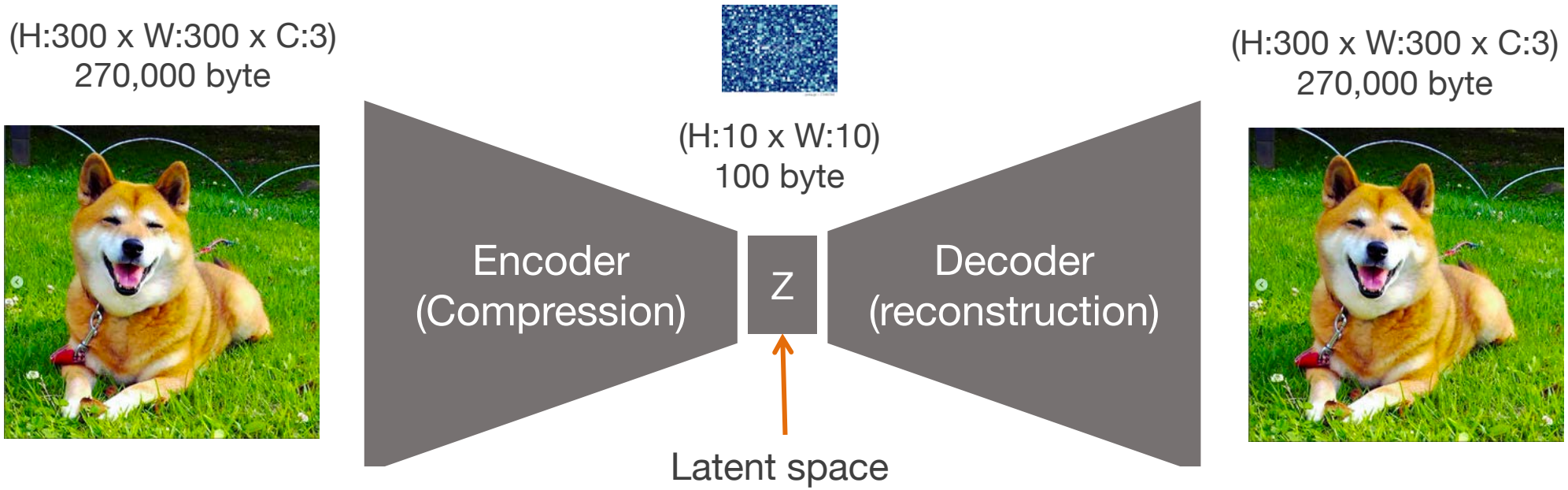
Qualia as Meta-Representations (a higher order theory with a novel interpretation of qualia)

Relationships between the input and output of neural networks are represented in the brain. Each coordinate in this embedding space (called qualia space) is a meta-representation of the corresponding neural network. The qualitative aspect of conscious experience (i.e. qualia) comes from the relationships of meta-representations represented in this qualia space.

Solution by Generation

What is a latent space?

- Low dimensional space where critical information is embedded
- Ex., In an autoencoder as below, latent space is created between the encoder and decoder.

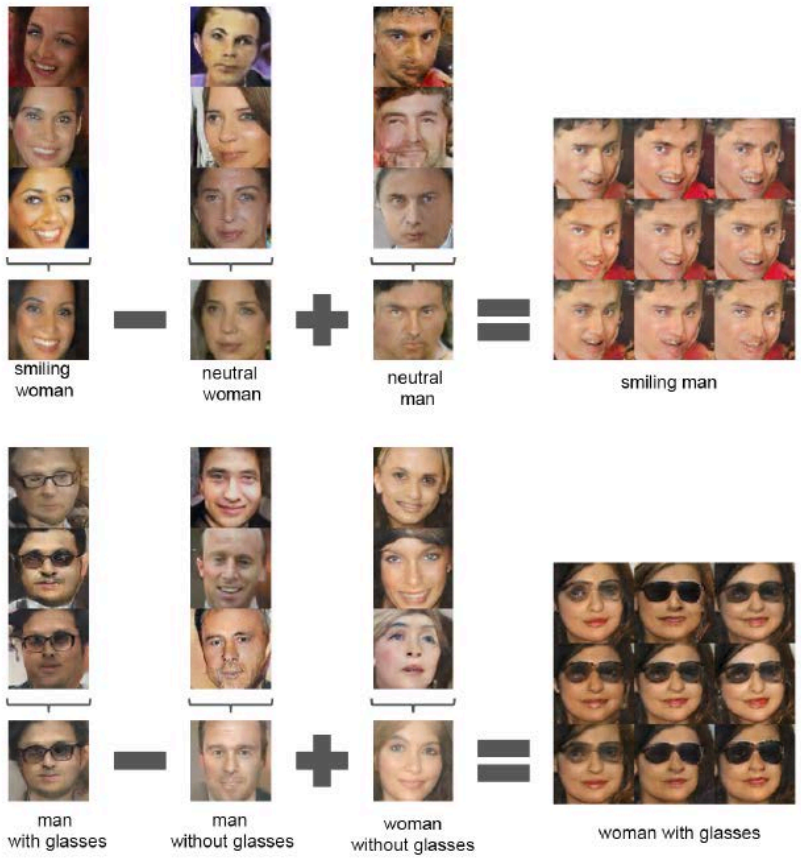


Solution by Generation

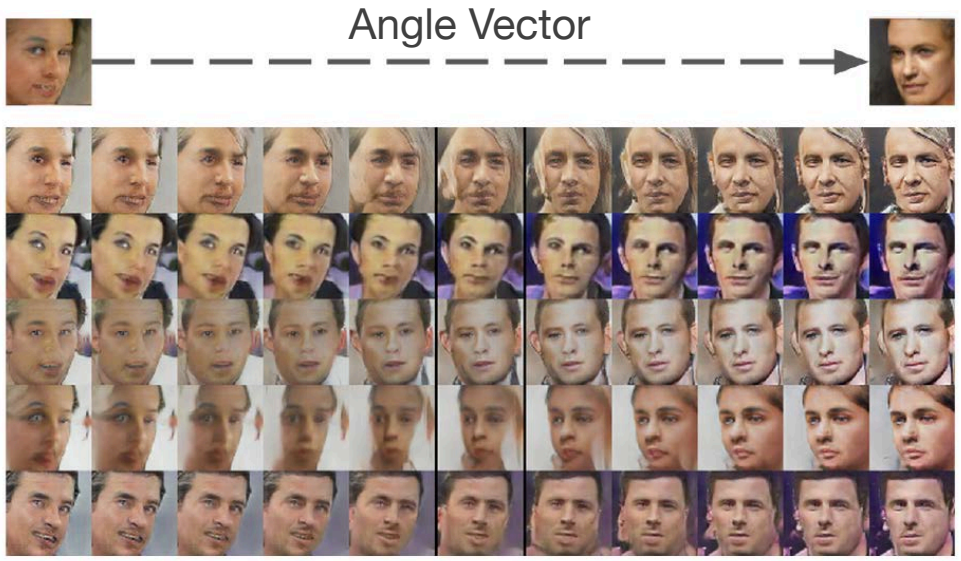
Embedding to latent space

- By embedding into latent space, we can perform arithmetic inside the latent space

Ex 1:
We can represent facial expressions, gender, the presence of glasses as vectors, and we can modify images by arithmetic's of those feature vectors.



Ex 2:
We can compute the turn vector using the average of a few leftward and rightward samples.
We can generate images as seen from different angles for any new faces.



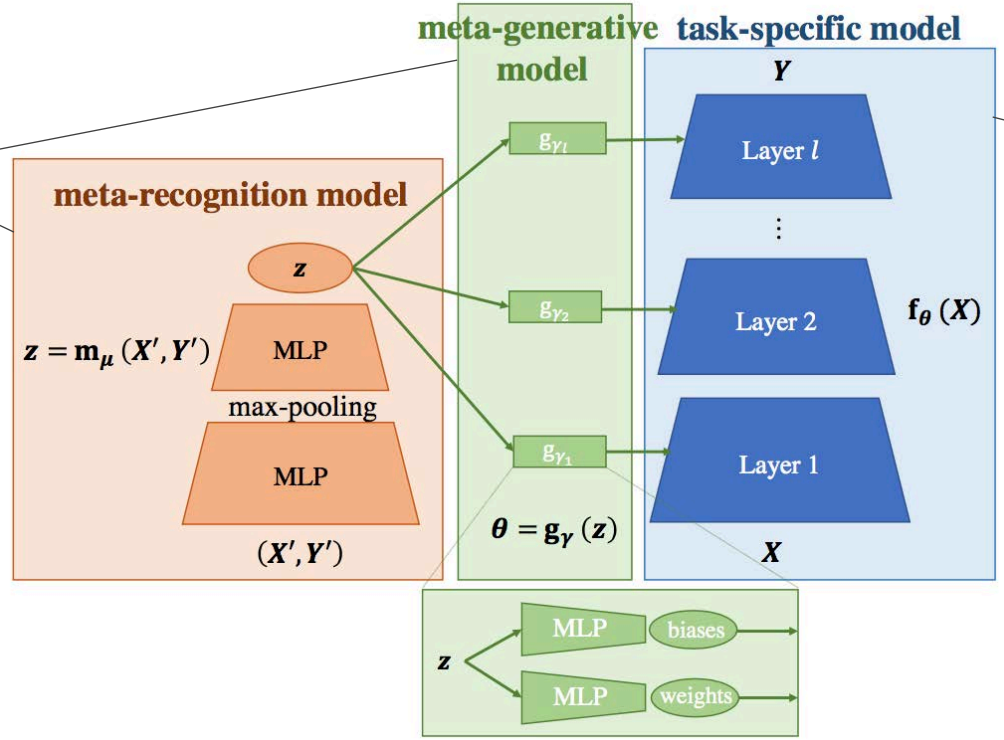
*) DCGAN https://arxiv.org/abs/1511.06434

Solution by Generation：MeLA

Meta-learning Autoencoders for few-shot prediction (Wu et al. 2018)

By training a metamodel with various task parameters, we can generate a new task specific model and solve a new task by using only a few-shot.

Meta-model to generate parameters of the task-specific model (blue)

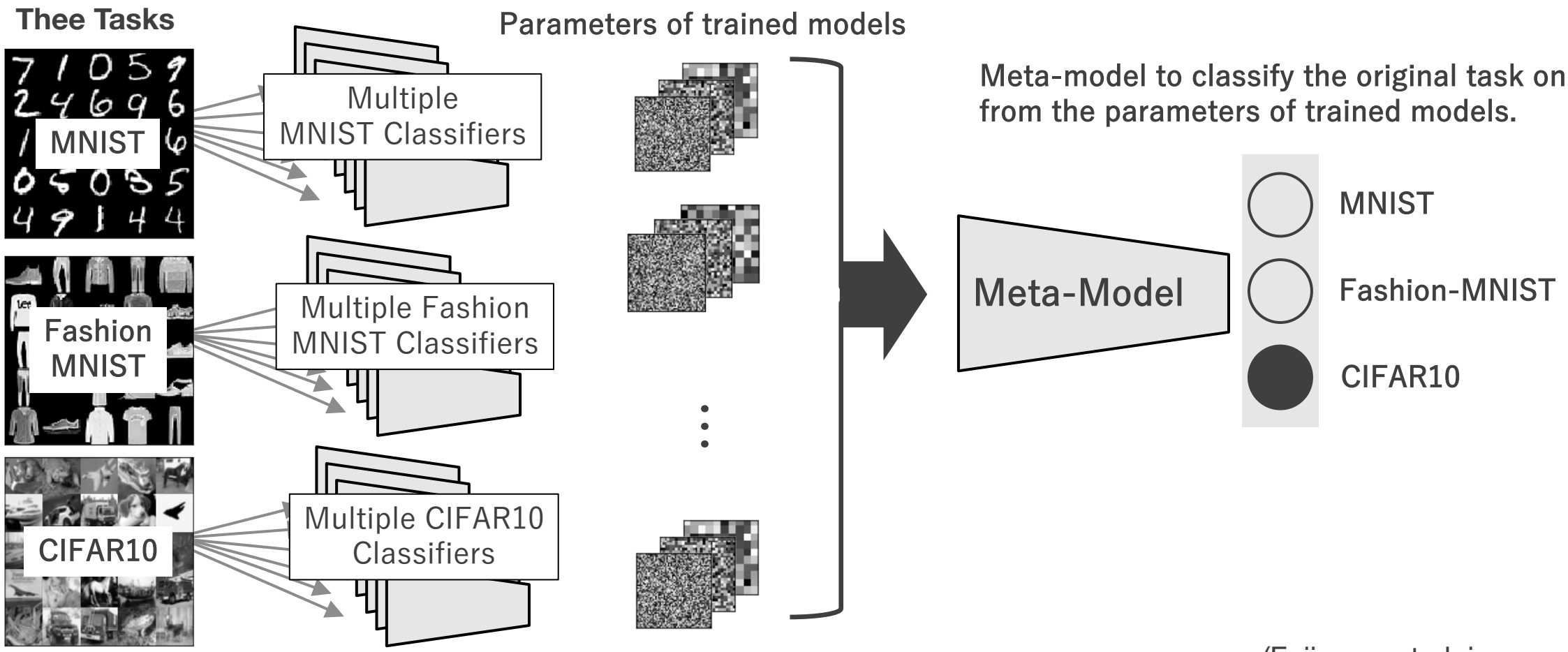


Task-specific model only solves a specific task.

(Wu et al. 2018)

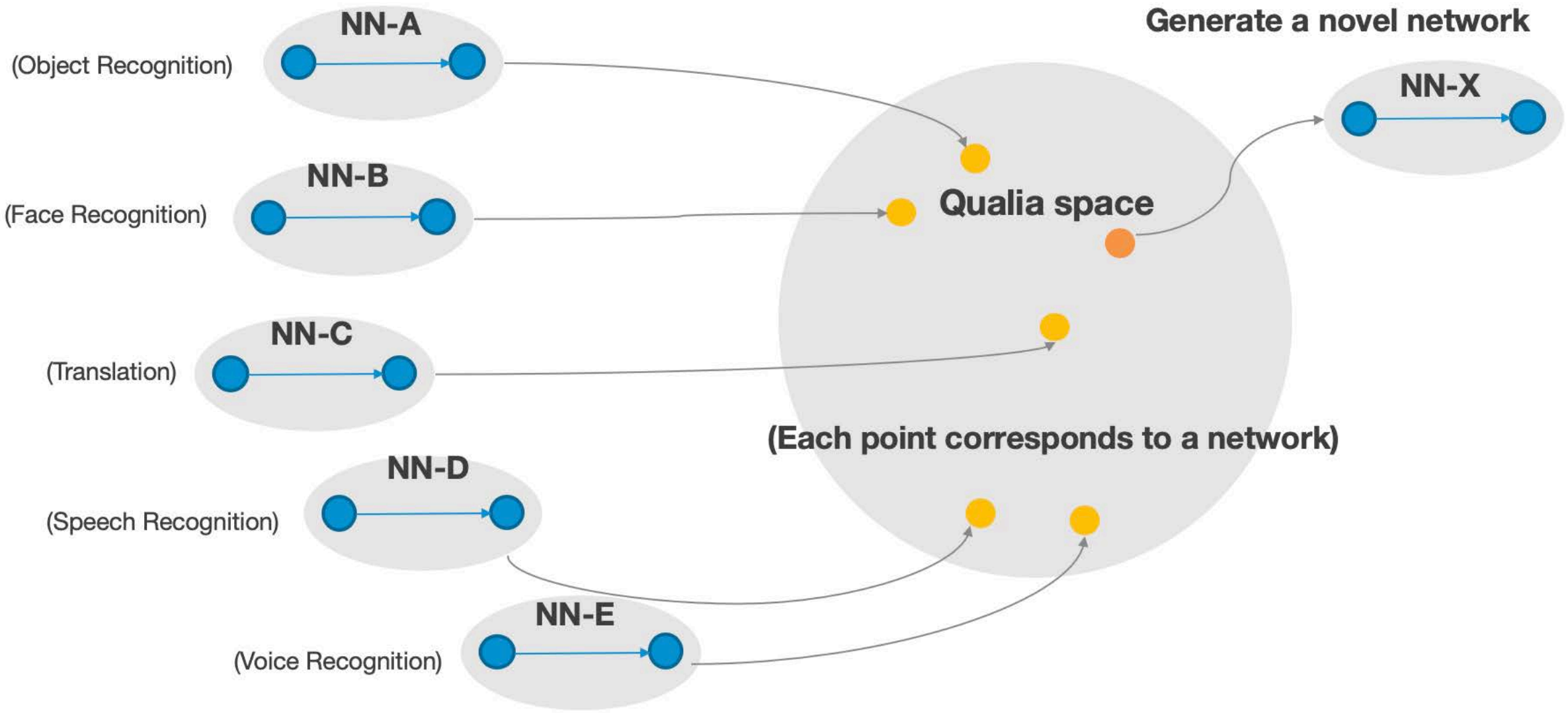
Meta Classifier Extracts Intrinsic Properties of Models

Meta-Discrimination Task: Distinguishing the original tasks from parameters of neural networks



(Fujisawa et al. in progress)

Embedding of Neural Networks and Qualia



Some thoughts on meta-representations

Thoughts

- Perhaps, the notion of embedding a network via their input-output relationships is in line with Yoneda's embedding (Tsuchiya & Saigo, 2020).
- Perhaps, generation of a network is not performed in the brain.
- But the meta-representation/qualia space is necessary for performing the solution by combination because it is important to have the ability to represent the quality of networks so that the agent knows which network can solve which task.
- Taking the embedding a meta-representation makes a lot of sense.
  - Our ability to analyze and compare different conscious experiences would require such a space.
  - It offers functional advantages of having such representations: we can use it for intuiting solutions for novel problems.
- This offers a more formal definition of meta-representation, whose notion has been vague.

Conclusions

Take home messages

- Three possible ways to build AGI proposed here have intimate connections with putative functions of consciousness.

- Considering those theories in terms of concrete implementations into AI clarifies relevant notions via more concrete computational processes (e.g. broadcasting as shared latent space, and meta-representation as embedding of functions).

- There might be a stronger link between consciousness and intelligence than previously thought. (Don’t confuse the two problems of the function of consciousness).

Ryota Kanai
kanair@araya.org
@kanair
Araya, Inc.

Special Thanks To My Collaborators:

Acer Chang (Araya, Inc.)
Martin Biehl (Araya, Inc.)
Yen Yu (Araya, Inc.)
Nicholas Guttenberg (ELSI/GoodAI)
Ippei Fujisawa (Araya, Inc.)
Masahiro Yasumoto (Araya, Inc.)
Hiro Hamada (Araya, Inc.)
Shinya Tamai (Araya, Inc.)
Atsushi Magata (Araya, Inc.)