Human-Machine
Harmonious
Collaboration

ARAYA

Two problems of Machine Consciousness

(and my random thoughts)

Ryota Kanai

CREST

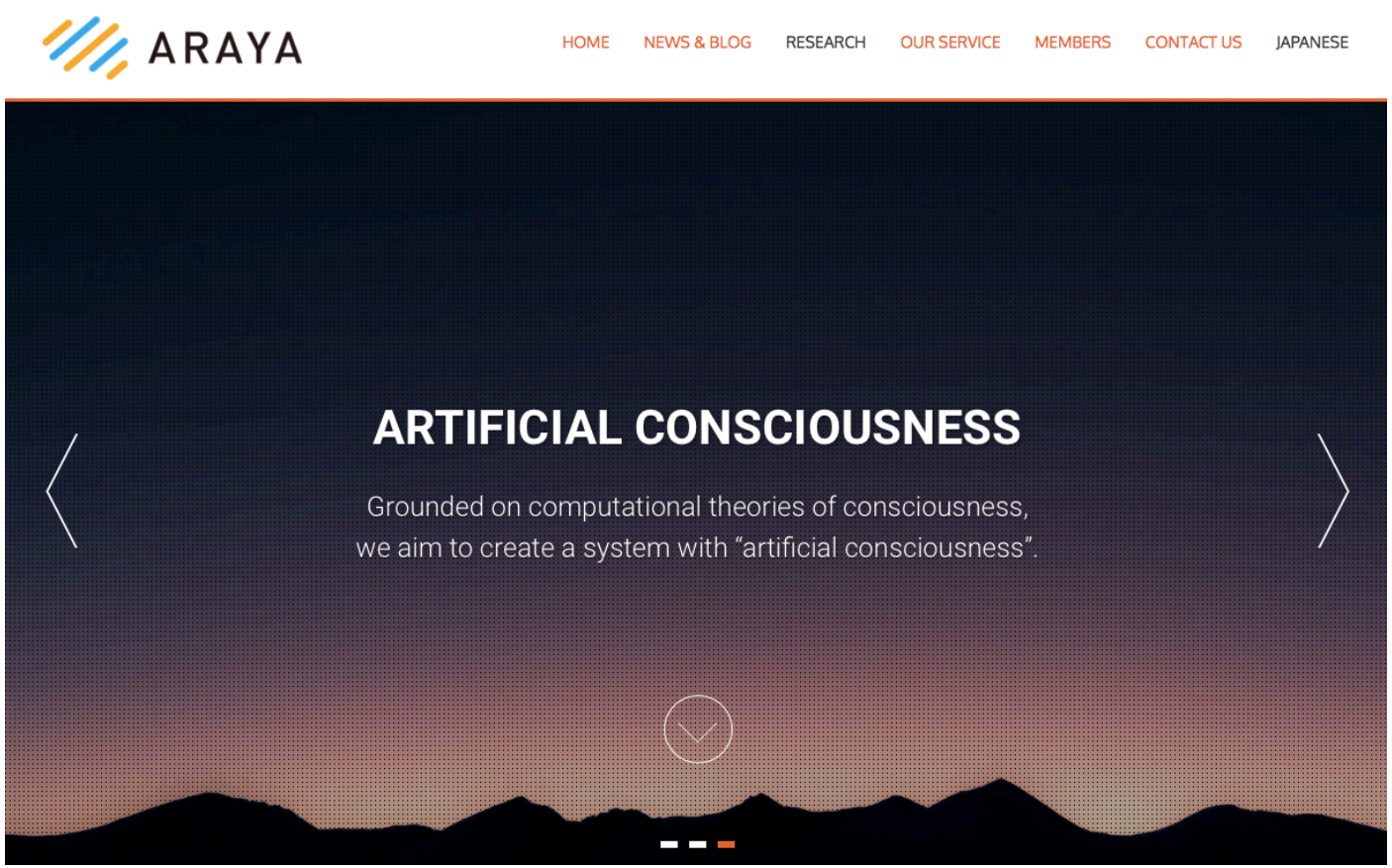
Araya, Inc.

Our Mission

We want to understand consciousness by making it.

Research

araya.org



Research Themes

- Integrated Information Theory (Oizumi, Kitazono)
- Intrinsic Motivation (Biehl, Magrans)
- Free Energy/ VAEs (Yu, Tamai)
- Coarse Graining (Chang)
- Meta-learning (Guttenberg)
- Wasserstein distance (Amari, Oizumi, Mizutani)

(We are trying to figure out new ways to do exciting research.)

YHouse:

yhousenyc.org

Consciousness Research Network

HOME CORN 2017 JOIN US CONTACT

conresnet.org

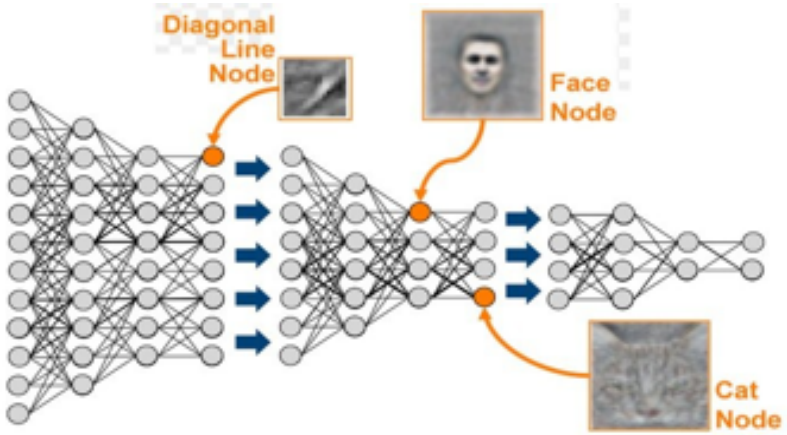
Why should we care about consciousness?

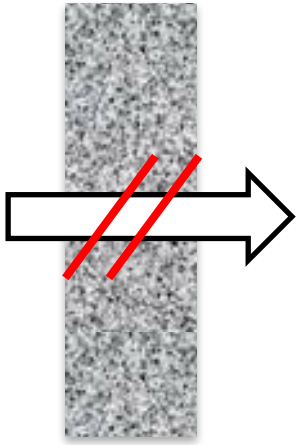
- **Consciousness is relevant for all kinds of cognition.**
  - Learning and Memory, Perception, Thoughts, Action, Decision Making, Emotion…

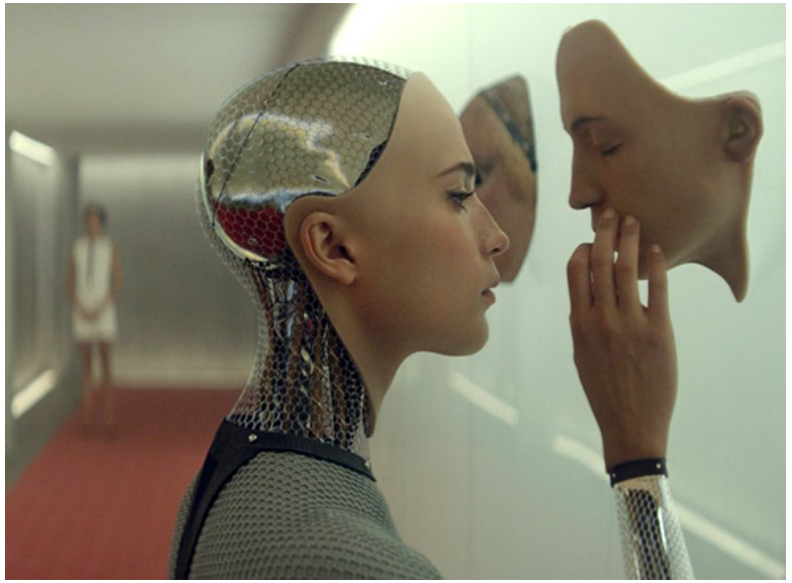
- **We don't understand intelligence unless we understand consciousness**
  - What does it mean to understand?
  - What does it mean to feel something?
  - What is consciousness for?

Current issues in AI

A huge gap from current AI to AGI







Current challenges in AI

1. Spontaneous behavior (e.g. Intention, Curiosity)
2. Generalization (e.g. Creativity, Understanding, Thought)
3. Explainability (e.g., Metacognition, Language)

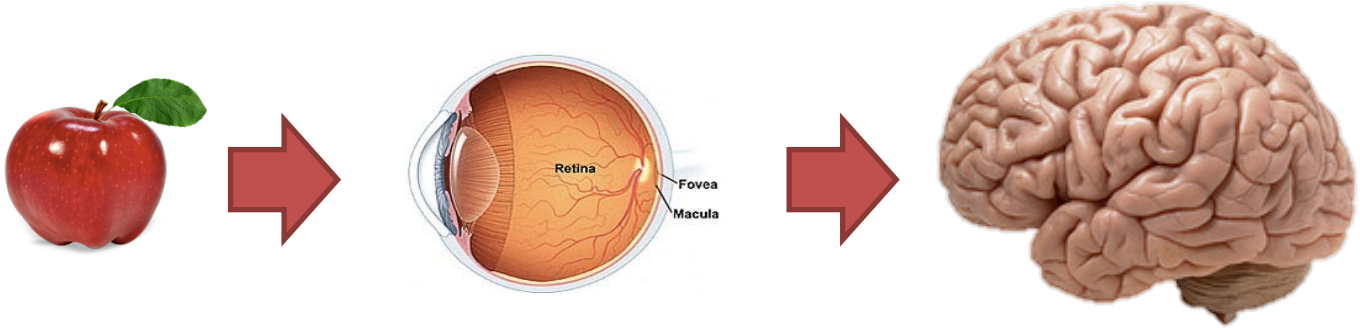
⇨ **We need to understand consciousness for AGI**

The problem of consciousness

How does the subjective experience (qualia) emerge from physical phenomena?

Subjective

Physical

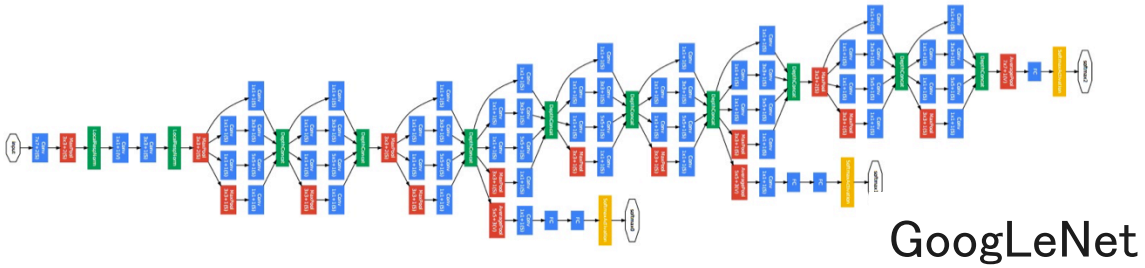


The problem of consciousness

Do current deep neural nets experience anything at all?

Subjective

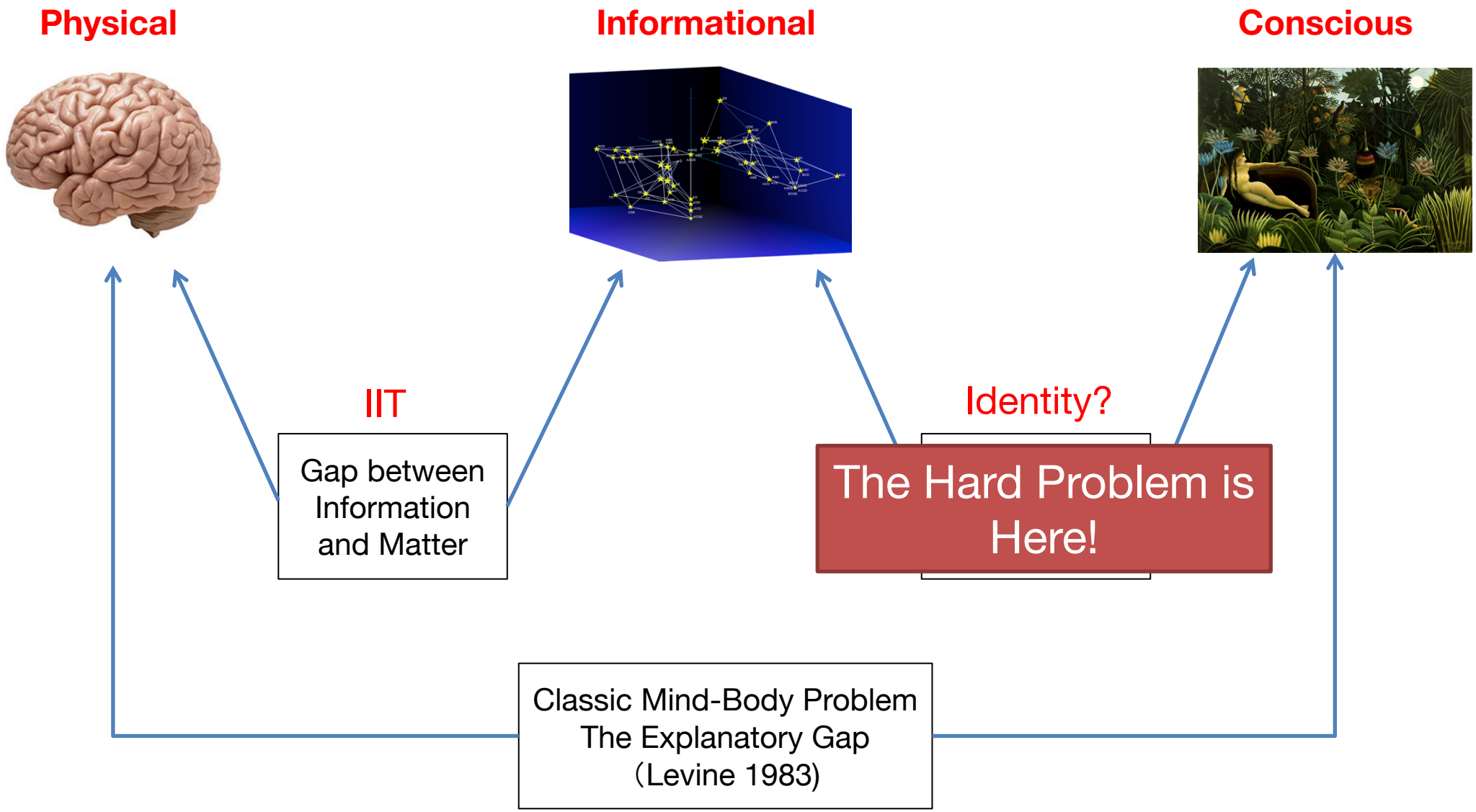
Physical



GoogLeNet

If not, why is it different from the brain?

The Hard Problem



We need a theory of consciousness

Bold ideas, unjustified anticipations, and speculative thought, are our only means for interpreting nature: our only organon, our only instrument, for grasping her. – Karl Popper

The Hard Problem will be dissolved by IIT 8.0 by 2030.

The Dual Aspect of Consciousness

Access Consciousness

- Objective and functional aspect of consciousness observed from the outside
- Reportability

Phenomenal Consciousness

- Subjective aspect of conscious experience only observed from the inside
- Qualia

The dual problems of machine consciousness

Creating Artificial Consciousness

- How can practically create a machine that has functions of consciousness?

Proving Artificial Consciousness

- How can we prove that it has phenomenal experience?

Four kinds of machine consciousness

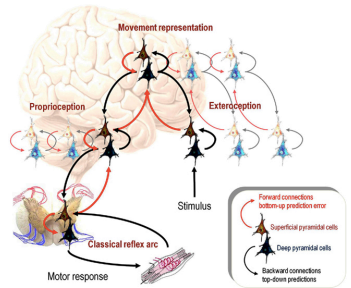
- MC1. Machines with the external behaviour associated with consciousness.
- MC2. Machines with the cognitive characteristics associated with consciousness.
- MC3. Machines with an architecture that is claimed to be a cause or correlate of human consciousness.
- MC4. Phenomenally conscious machines.

Gamez (2008)

What we do at Araya

Creating AC

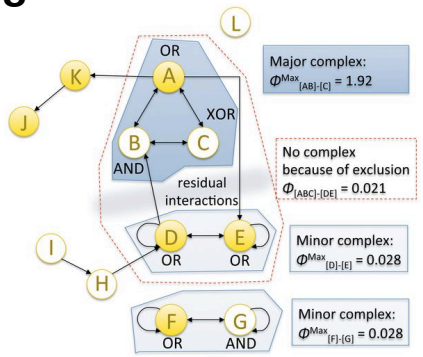
Free Energy Principle



Maximise $F(s, \mu)$

Proving AC

Integrated Information Theory

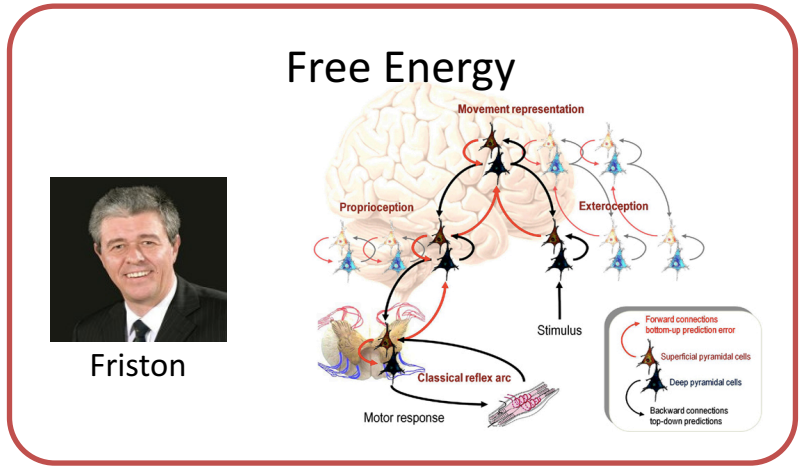


Maximise Φ

Emerging theories

Free Energy Principle

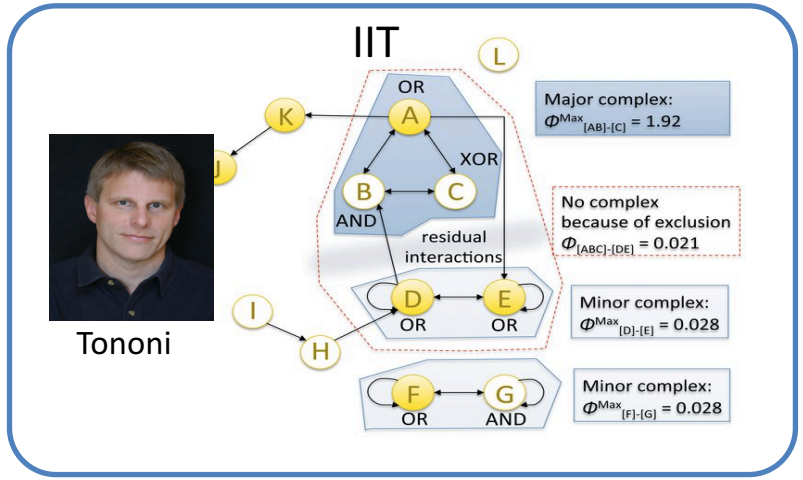
1. General computational mechanism of the brain
2. Includes action selection, agency, counterfactuals.
3. Easier to construct function



Integrated Information Theory (IIT)

1. Axioms from phenomenology
2. Emphasis on intrinsic perspective
3. Identity claim

   - an experience is identical to the maximally irreducible conceptual structure (MICS)



Measuring Consciousness

Measuring Machine Consciousness

- Key properties for a consciousness measure
  - Intrinsic Perspective of information structure in a system
  - Axiomatic Theory starting from phenomenology
  - Identity between subjective experience and information structure

Integrated Information Theory (IIT) satisfies these conditions.

Intrinsic Perspective

Shannon's Information Theory
(Extrinsic Perspective)

Stimulus

Neural Activity

Correspondences are assigned by the experimenter

Labels are given by an external observer

Integrated Information
Theory
(Intrinsic Perspective)

Brain can see only itself

No labels

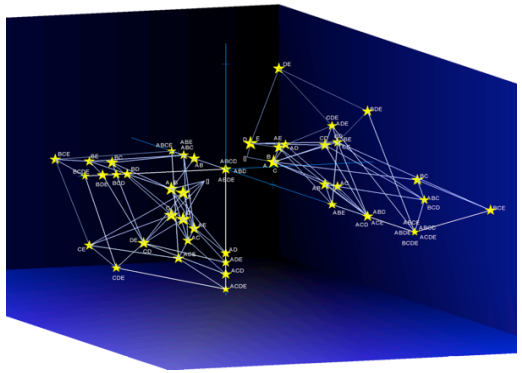
Axioms

- **Existence**: Consciousness exists – it is an undeniable aspect of reality.
- **Composition:** Consciousness is compositional (structured): each experience consists of multiple aspects in various combinations.
- **Information**: Consciousness is informative: each experience differs in its particular way from other possible experiences
- **Integration**: Consciousness is integrated: each experience is (strongly) irreducible to non-interdependent components.
- **Exclusion:** Consciousness is definite and exclusive: each experience has a particular spatial and temporal grain.

Identity

- Identity: there is an identity between phenomenological properties of experience and informational/causal properties of physical systems

Intrinsic Information Structure



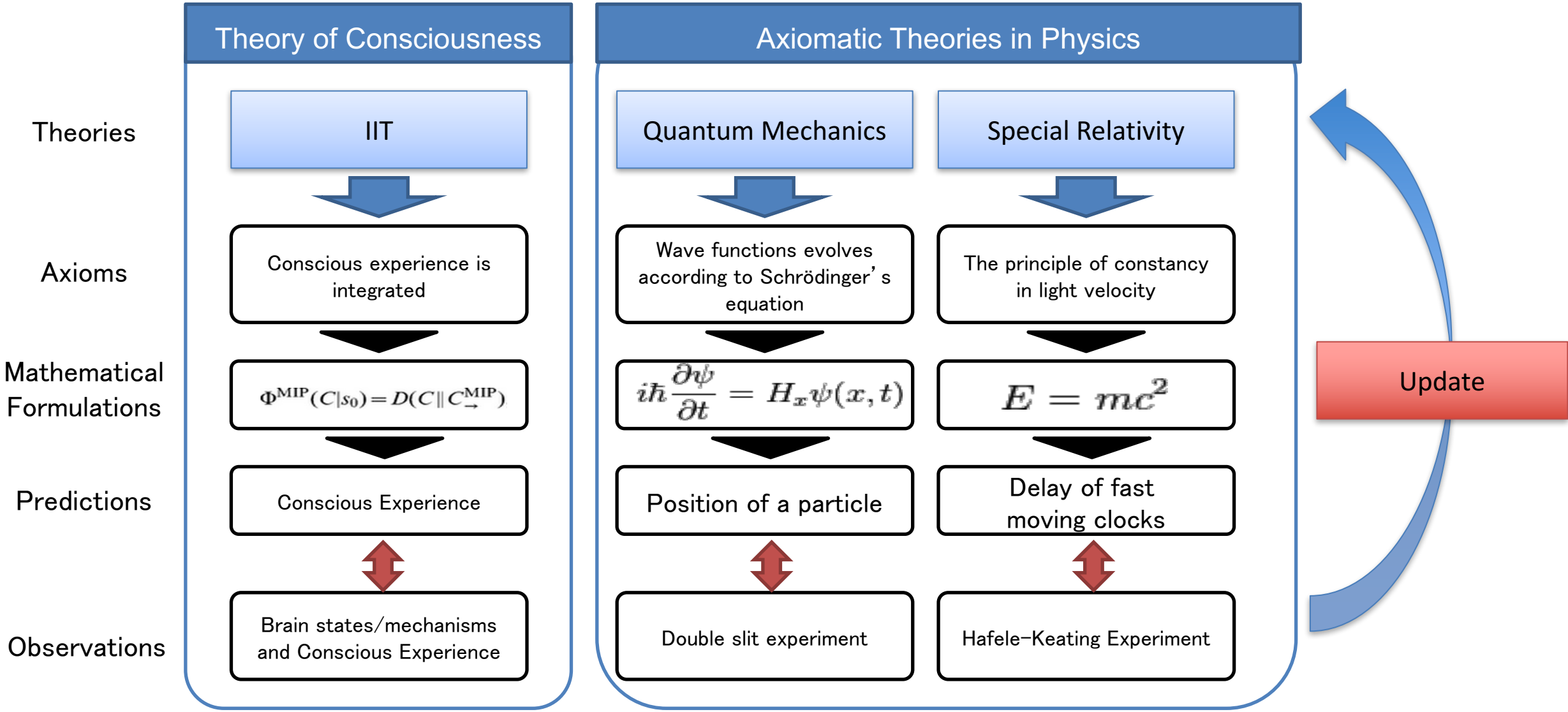
MICS

$$\Phi = \{$$

Conscious Experience

$$\}$$

Axiomatic Theories in Science

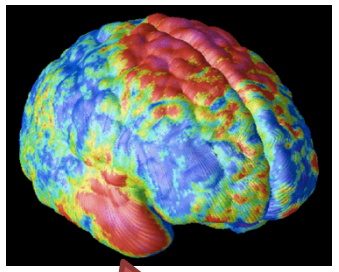


Axiomatic approach

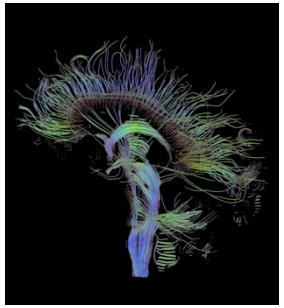
Axioms: Top-down approach
(Tradition of physics)

Consciousness

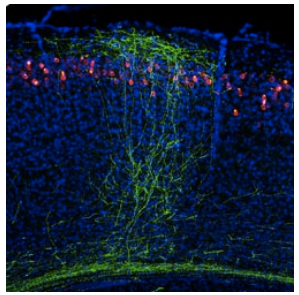
Whole Brain



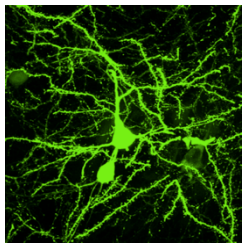
Global Network



Local Circuit



Neuron



The Hard Problem

Explanation

Explanation

Explanation

Reductionism: Bottom-up
(Biology)

Discussion point 1: Axioms

- Compare axioms of IIT with axioms in mathematics, physics, and other fields
  - Physics
    - Newton's laws of motion
    - Axioms of quantum mechanics
  - Mathematics
    - Euclidian geometry
    - Kolmogorov's axioms of probability
    - Peano's axioms of natural numbers

Discussion point 1: Axioms

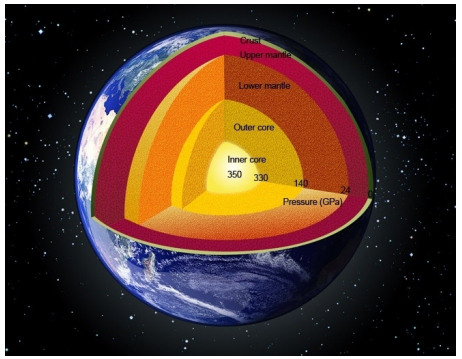
- What are the key properties of successful axioms?
- Are the axioms of IIT phenomenologically trivial?
- Should there be other axioms?
- Is phi uniquely defined?

Inferences about unobservables

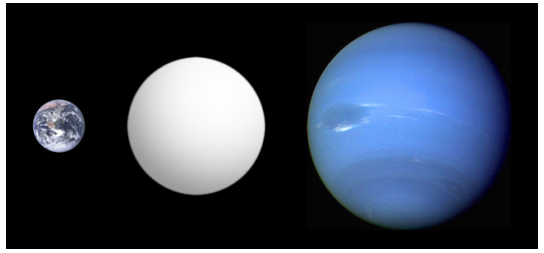
In science, we deal with many things that are not directly observable.

Seasons in Australia

The Core of the Earth



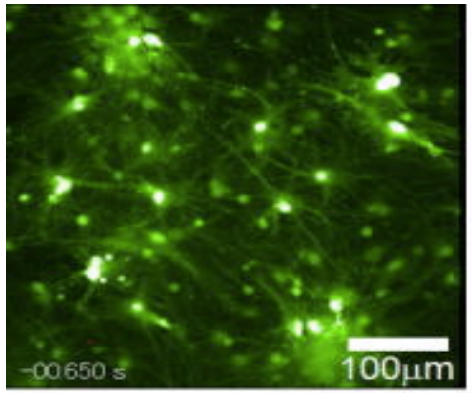
Exoplanets



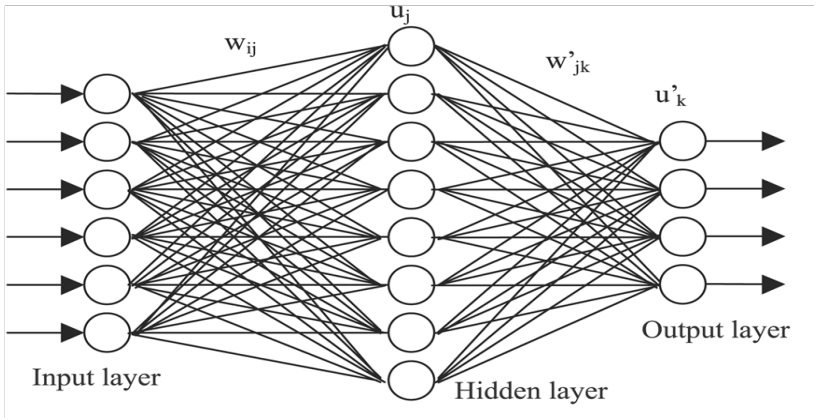
We understand nature through the lens of theories.

Inferences about Artificial Consciousness

Neurons on a petri dish



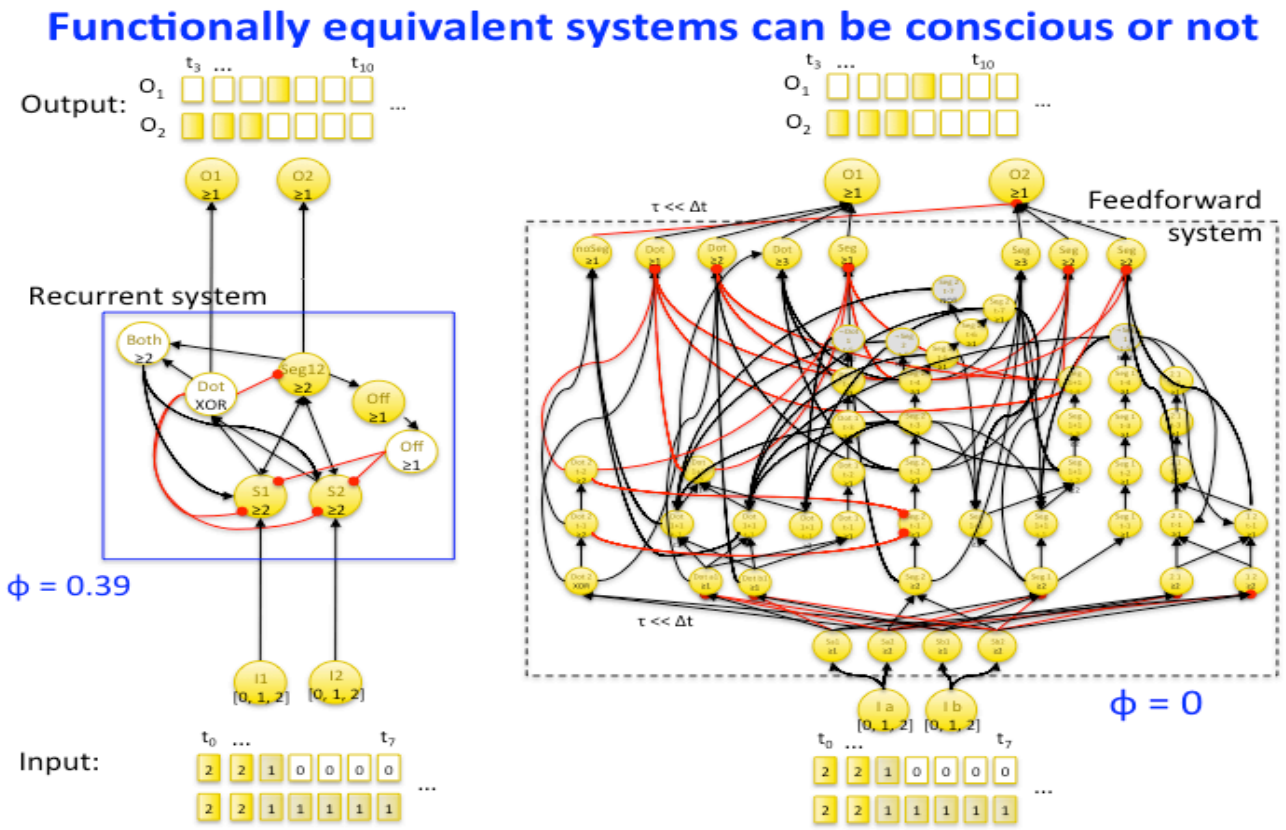
Artificial Neural Network



Computer programs in a robot

Are they conscious?

Some simple cases

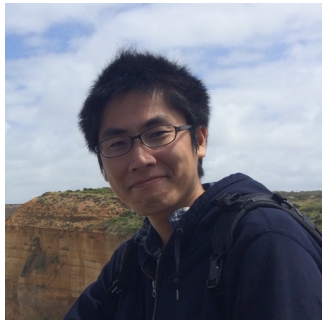


We can quantify the degree of consciousness in artificial systems.
Identical functions do not imply the same consciousness.

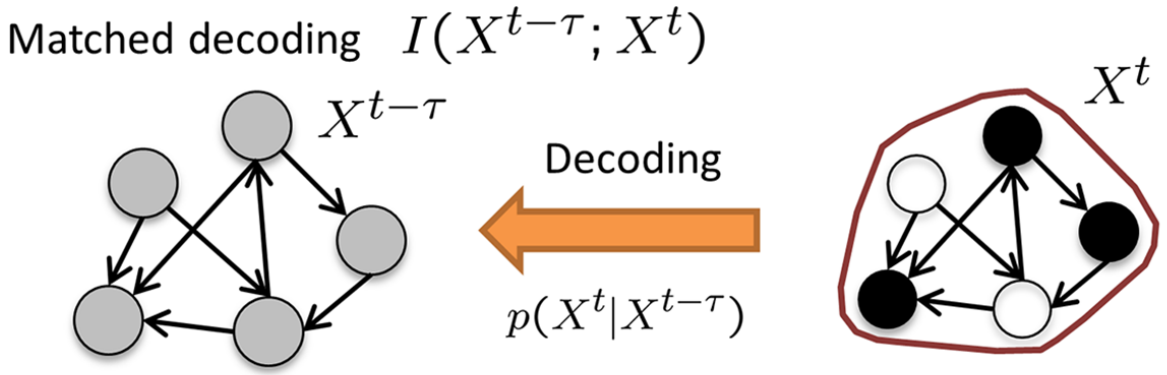
Practical Issues with IIT

- We cannot identify all causal relationships in real nervous systems.
- The number of possible partitions explodes (= a difficulty in computation).
- It is required to observe the all nodes at the same time (= a difficulty in data acquisition).

Measuring phi from empirical data



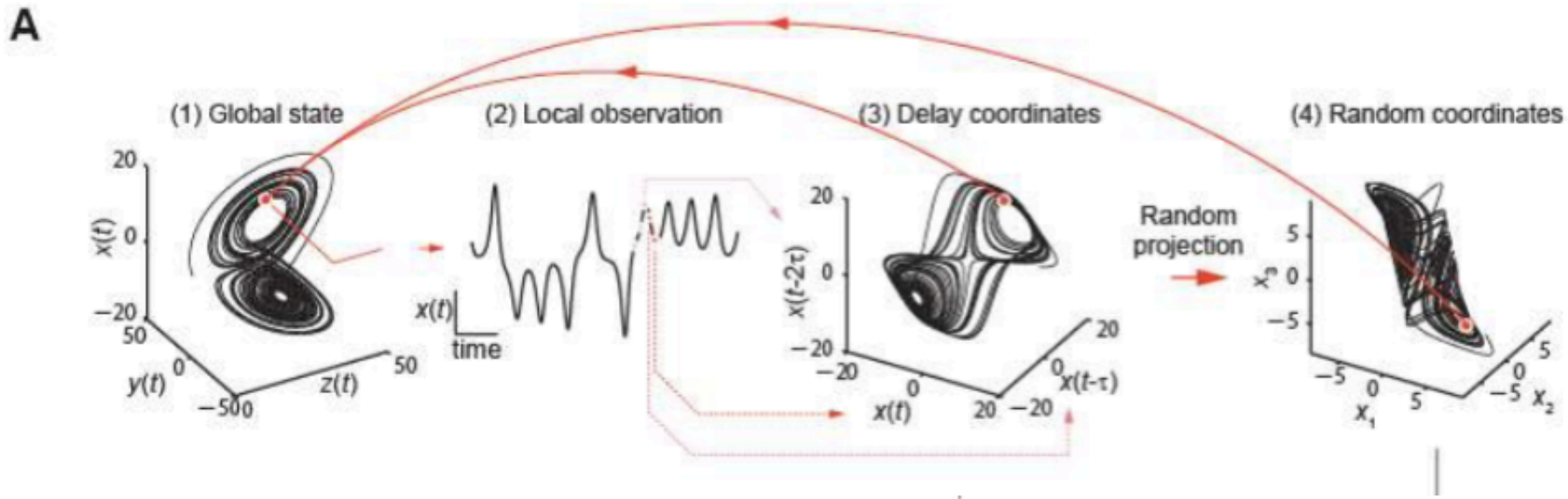
Masafumi Oizumi

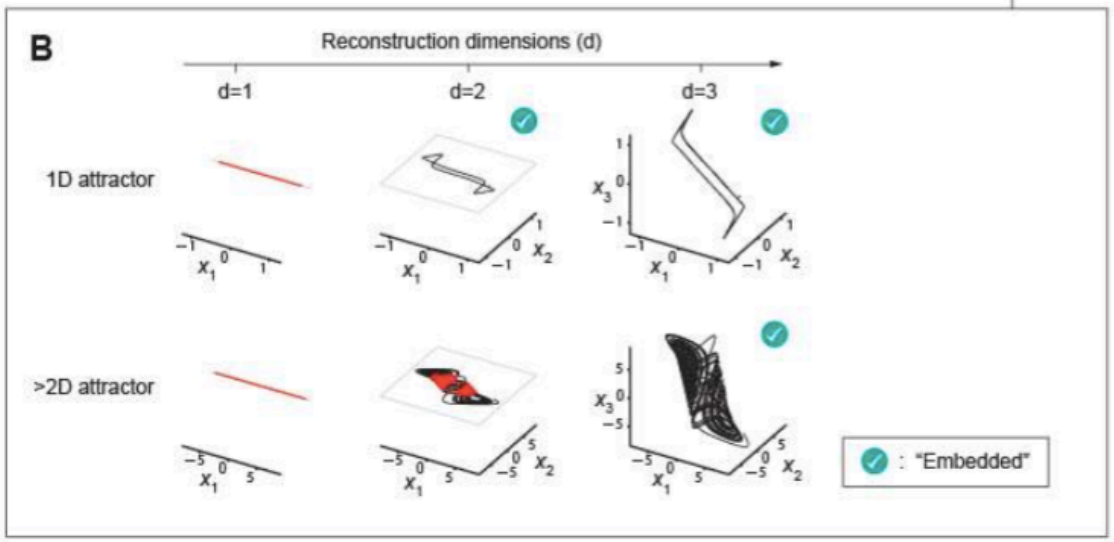


Oizumi et al. (2016) PLoS Comp

Integrated information in continuous attractors

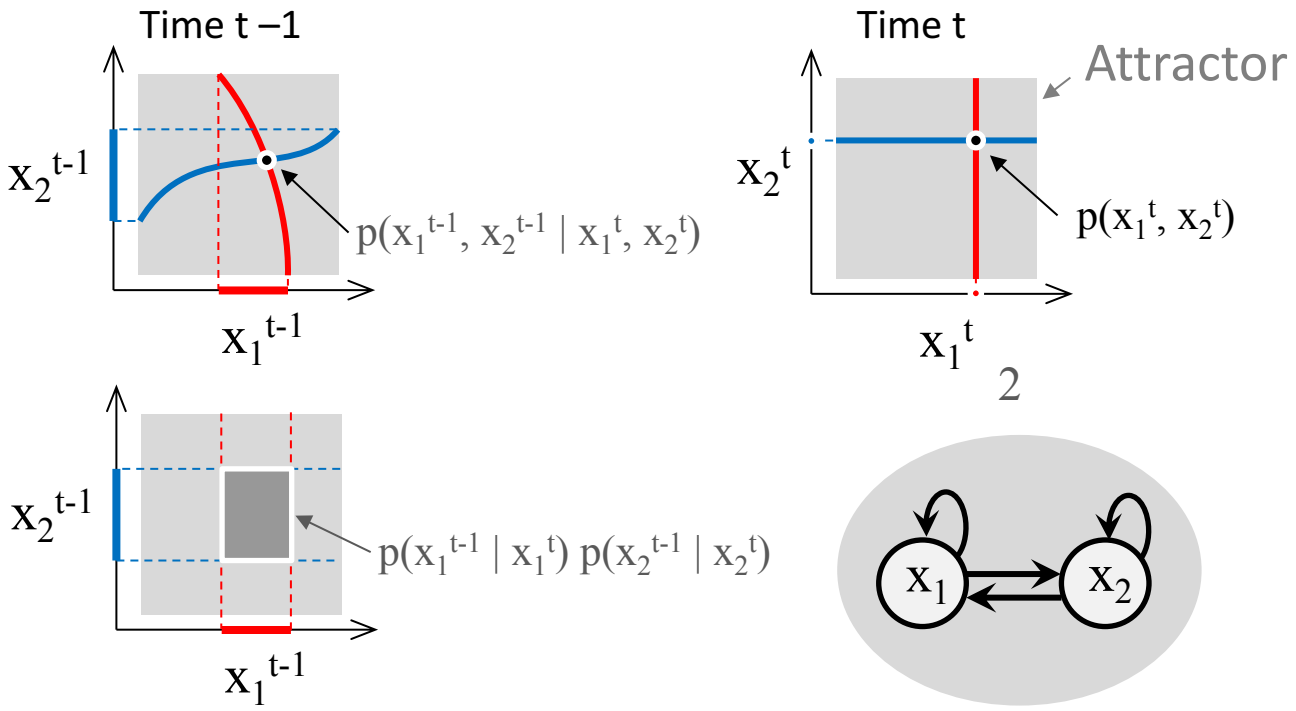
Satohiro Tajima
(Univeresité de Genève)





Tajima et al. (2015)

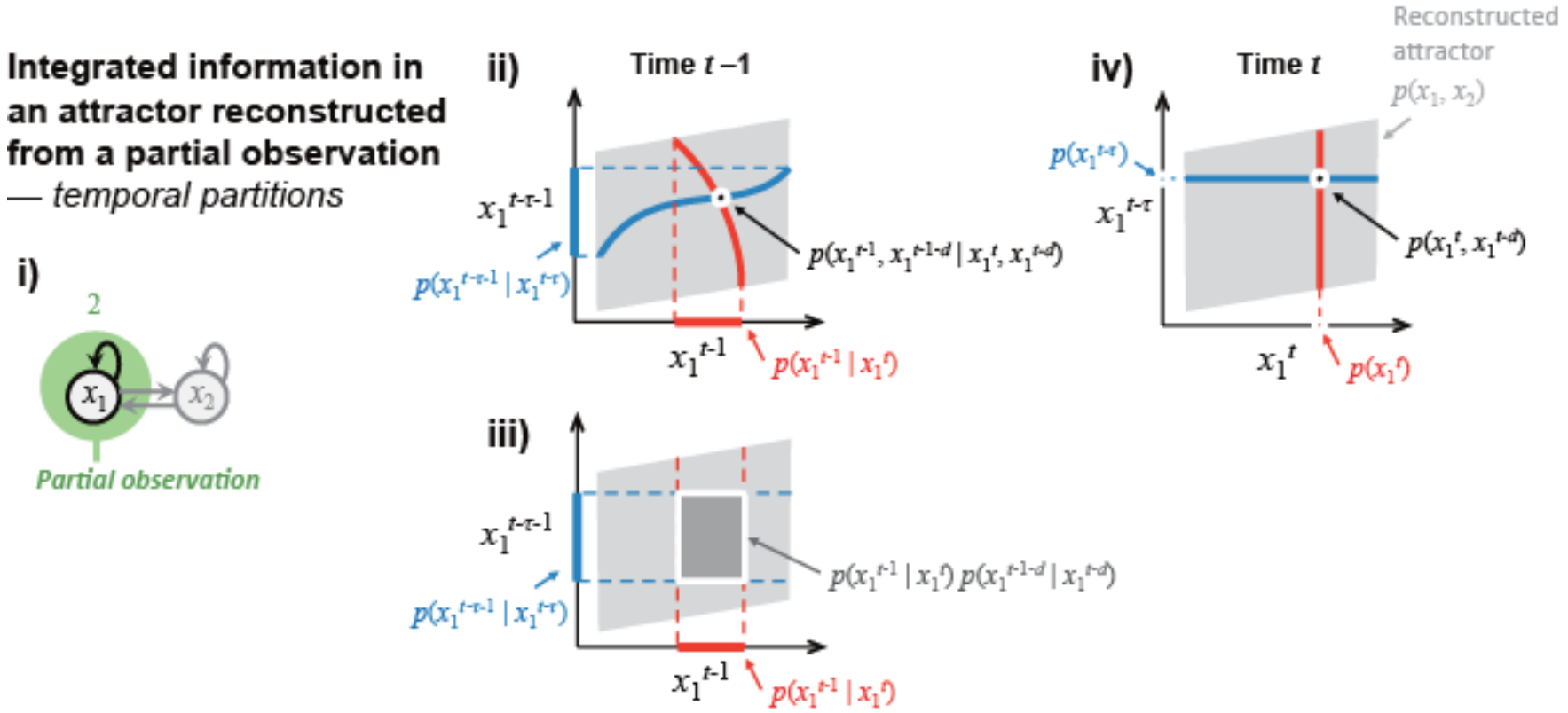
Integrated information in continuous attractors



$$\varphi^{\mathrm{Dim}} \equiv \mathrm{Dim}[p(x_1^{t-1}|x_1^t)p(x_2^{t-1}|x_2^t)] - \mathrm{Dim}[p(x_1^{t-1}, x_2^{t-1}|x_1^t, x_2^t)].$$

Tajima & Kanai (2017)

Integrated information in continuous attractors

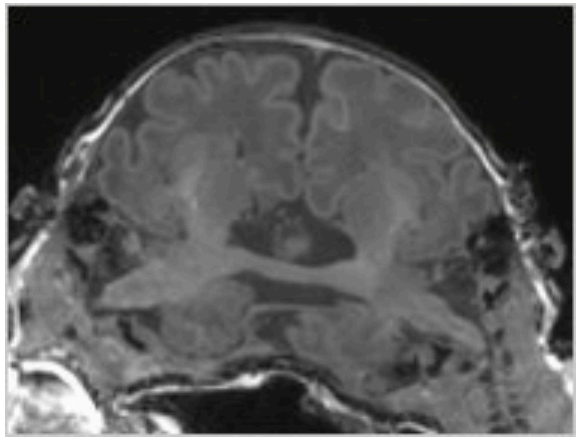


- We can recover the topology of attractor and estimate the dimensionality from partial observation.
- We speculate quality may be captured by topological characteristics.

Tajima & Kanai (2017)

Ultimate method: Direct connection

The Hogan sisters seem to share qualia via direct connections in the thalamus.



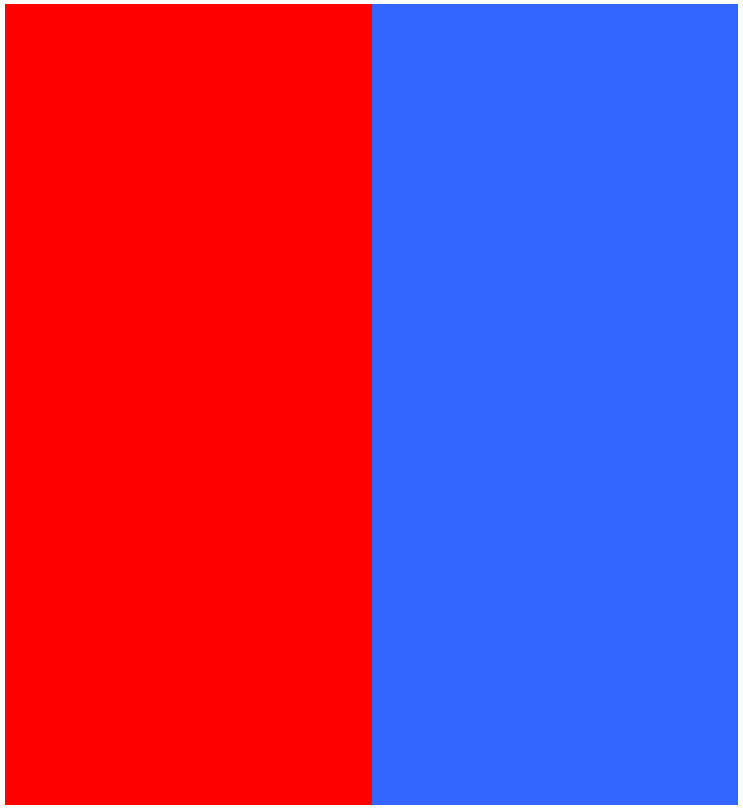
- Qualia seem to be shared among the brains through direct connections.
- We can connect brains directly to AI to check the qualia in AI.

Functions of Consciousness

We need to wake up from Chalmer’s epiphenomenal dogma.

The Dual Aspects of Consciousness

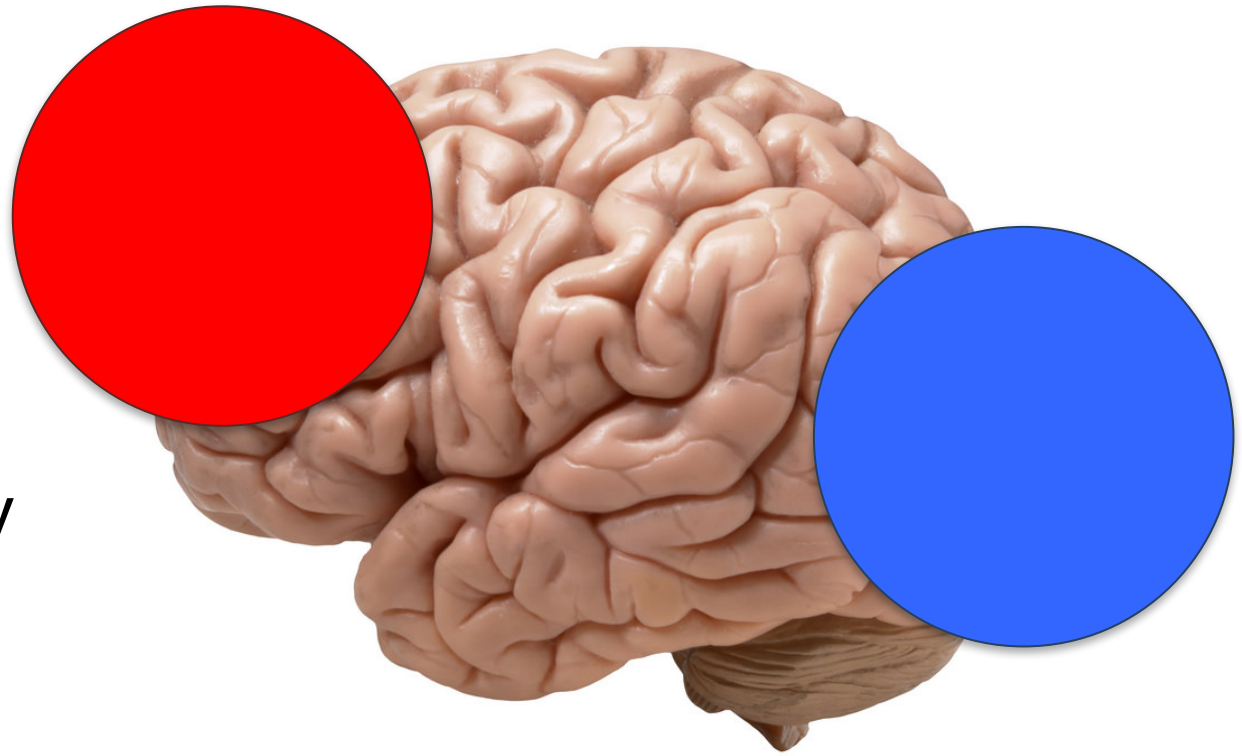
One Consciousness with Dual Aspects



Objective	**Subjective**
Access	Phenomenal
Public	Private
Views from outside	Views from within

Sensation and Action in Consciousness

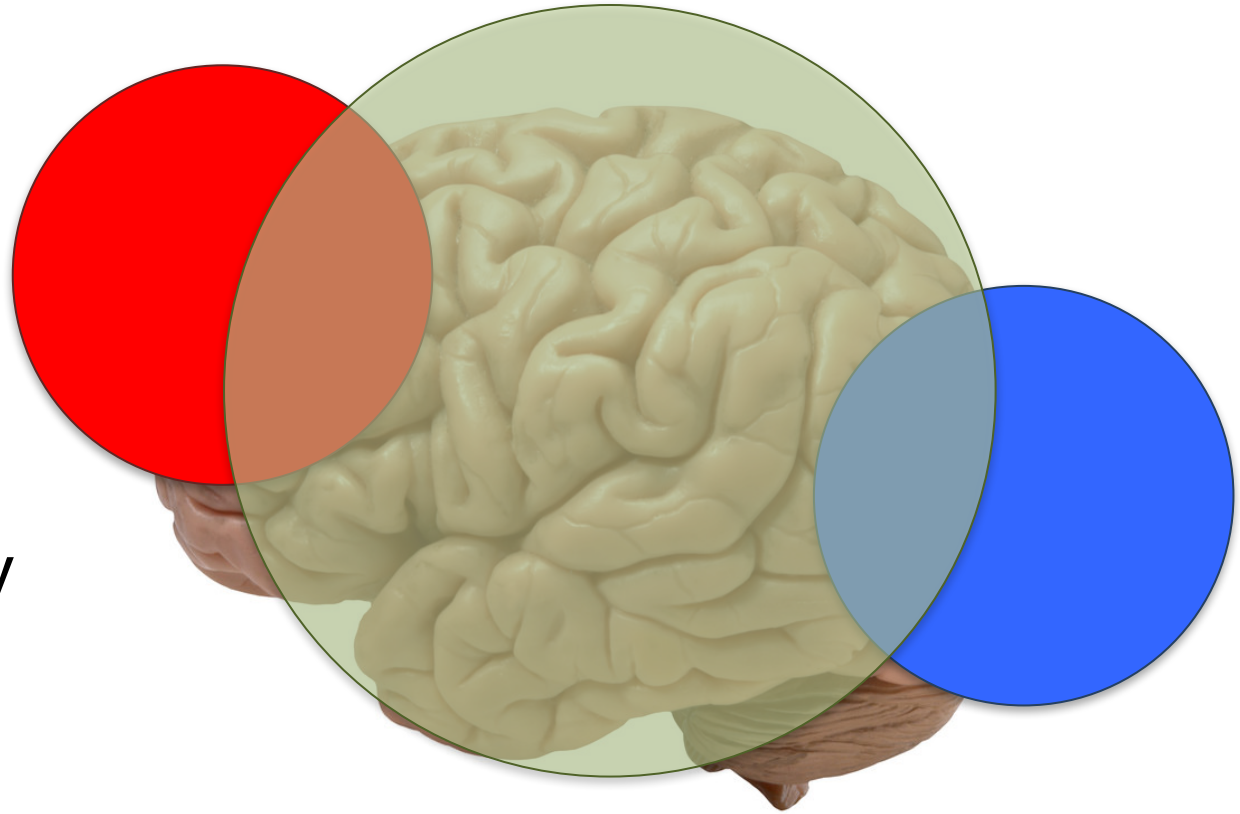
Access
Planning
Executive
Free Will
Report
Working Memory
Thought



Phenomenal
Sensation
Qualia
Ineffable
Feeling

Sensation and Action in Consciousness

Access
Planning
Executive
Free Will
Report
Working Memory
Thought



Phenomenal
Sensation
Qualia
Ineffable
Feeling

I want to abolish this dichotomy

Searle's Biological Naturalism

- Consciousness is causally reducible to brain processes.
  - All the features of consciousness are accounted for causally by neurobiological processes.
- But this does not imply ontological reduction.
  - First person ontology versus third person ontology

Rationale for creating consciousness

Biological Naturalism suggests:

- If we create the causal structure of the third-person aspect of consciousness, the first-person aspect of consciousness should be generated. (MC3→MC4)

- The problem of consciousness is to find the third-person observable causal structure of consciousness.

- Identification of such functional aspects is the first step.

Tinbergen's four questions

Four kinds of why biological systems have particular functions as applied to consciousness.

- **Function（adaptation）**
  - What is the biological advantage (Artificial consciousness)
- **Phylogeny (evolution)**
  - How did consciousness emerge in the course of evolution?
  - Which species has consciousness?
- **Ontogeny (development)**
  - At what stage of brain development does consciousness appear?
  - How does experience shape the contents of consciousness?
- **Mechanism (causation)**
  - What is the mechanism that generates consciousness?
  - What is the minimal set of necessary and sufficient condition for consciousness?（NCC，MICS）

Functions of Consciousness?

- Perception: Ability to discriminate sensory stimuli
- Short Term Memory: The ability to hold sensory information for report
- Metacognition: The ability to report the quality of sensory experience such as vividness or confidence
- Attention: the ability to focus on selected aspects of sensory stimuli
- Executive Control: the ability to control behaviour according to a current goal
- Symbolic Operation: the ability to manipulate symbols

Discussion point 2: Functions of consciousness

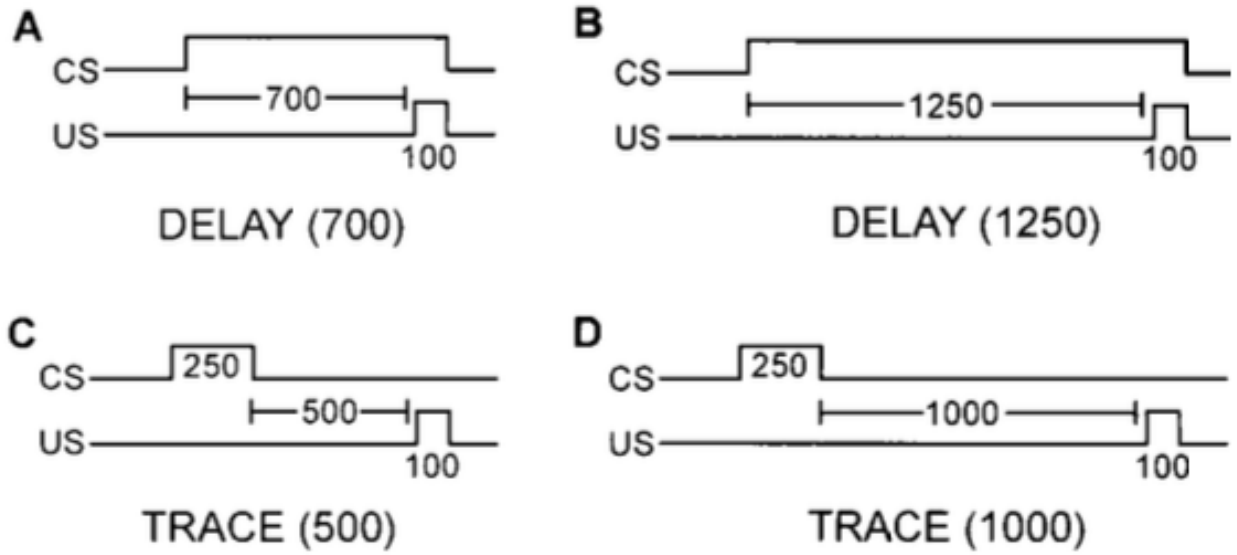
- What are the functions of consciousness?
- Is there anything we can’t do without consciousness?
- Is any cognitive function sufficient for proving consciousness?

Consciousness is
for bridging a temporal gap

- Trace conditioning.

- Patient with visual form agnosia (DF).

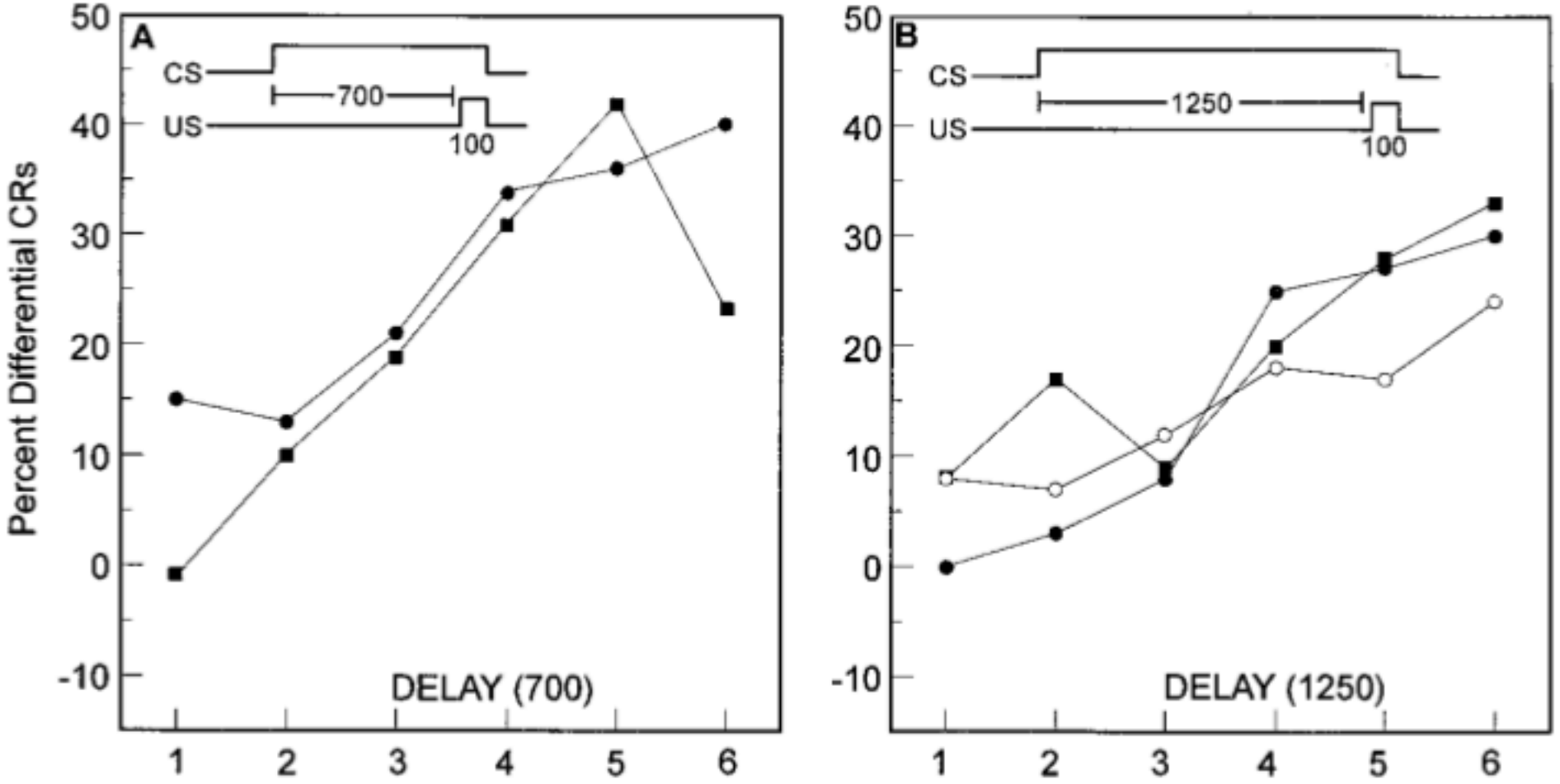
Trace conditioning



- Delay conditioning occurs without awareness
- Trace conditioning requires awareness (and hippocampus)

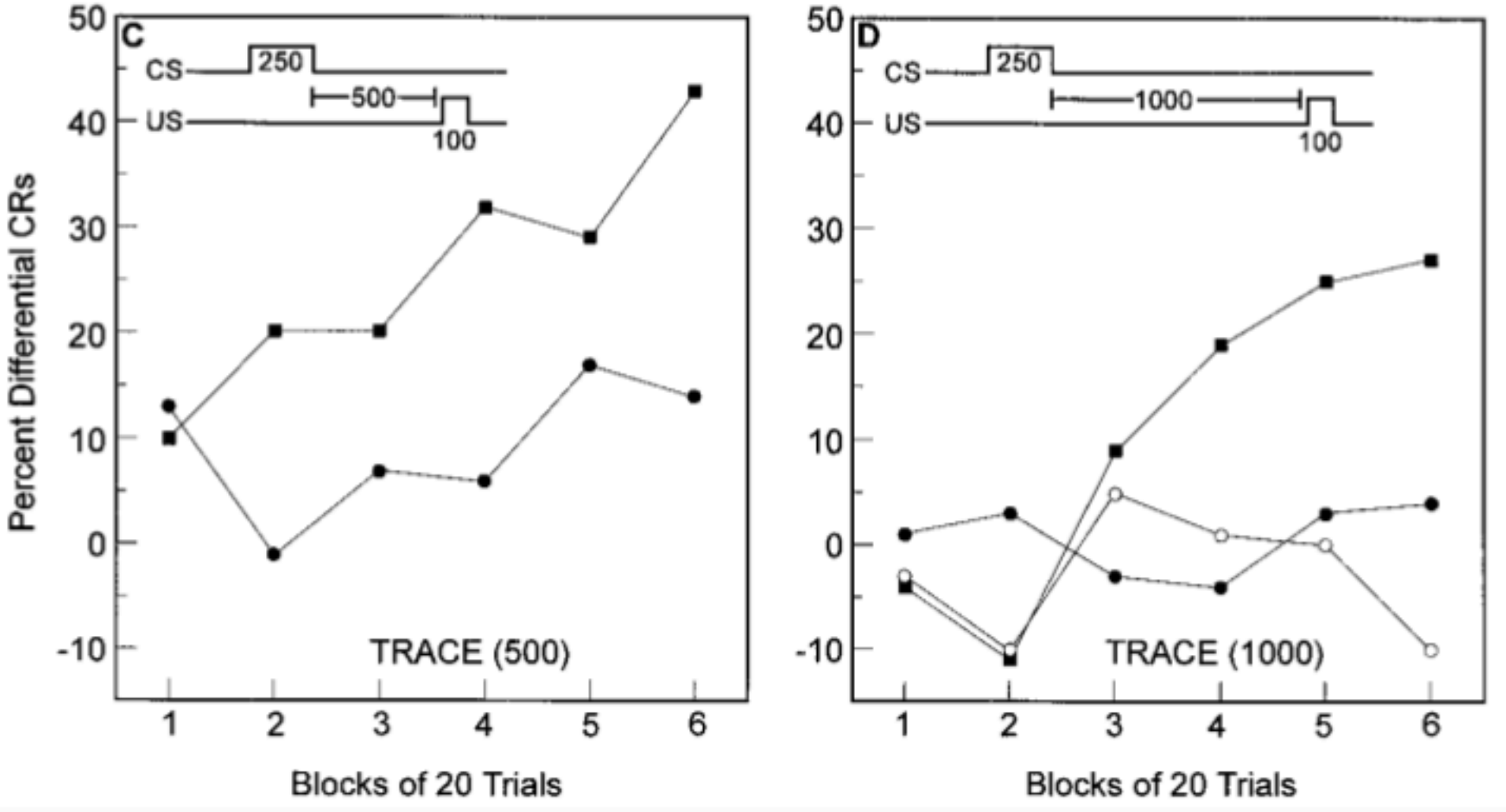
Clark & Squire (1998)

Trace conditioning



Clark & Squire (1998)

Trace conditioning

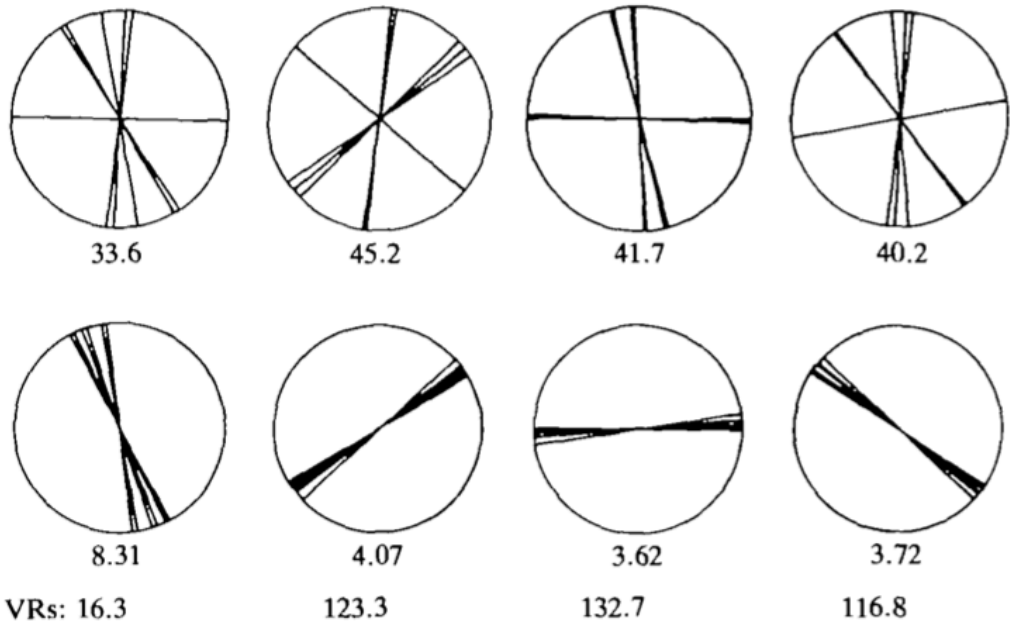


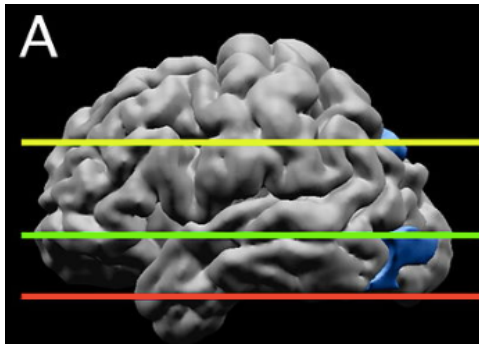
Clark & Squire (1998)

Online and memory guided action

Perceptual judgement

Visually guided action





Memory guided action depends on perceptual system

Milner et al. (1991)

The first clue

A potential function of consciousness might be broadening of temporal window on the world – to give the present moment an extended duration.

V1 hypothesis

- V1 is not conscious because it does not project to prefrontal cortex.

- The rationale from biological usefulness of awareness
  - A: To produce the best current interpretation of the visual scene
  - B: To make the information available for the system that plan and execute voluntary motor outputs.

Crick & Koch (1995)

Back to Ramachandran's Laws

1. Irrevocable
   - Qualia are protected and insulated from top-down influences.

2. open-ended/flexible
3. short-term memory
4. Attention

Corresponding Dennett's Creatures

- Darwinian creatures: Simple selection and existence
- Skinnerian creatures: Simple Associative Learning
- Popperian creatures: Learning from Counterfactuals
- Gregorian creatures: Learning and culture

A Hypothesis on Consciousness

A Counterfactual Information Generation Hypothesis

Agents with the ability to generate counterfactual predictions of their own state in the environment have consciousness.

Counterfactuals

ARAYA
BRAIN IMAGING

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

A model of the world, and sensory motor contingency are prerequisites for the ability to simulate the world internally.

Functions achieved

- Intention/planning
- Imagination/thought
- Retrospection/short-term memory

Aleksander's Axioms of MC

- Depiction. The system has perceptual states that 'represent' elements of the world and their location.
- Imagination. The system can recall parts of the world or create sensations that are like parts of the world.
- Attention. The system is capable of selecting which parts of the world to depict or imagine.
- Planning. The system has control over sequences of states to plan actions.
- Emotion. The system has affective states that evaluate planned actions and determine the ensuing action.

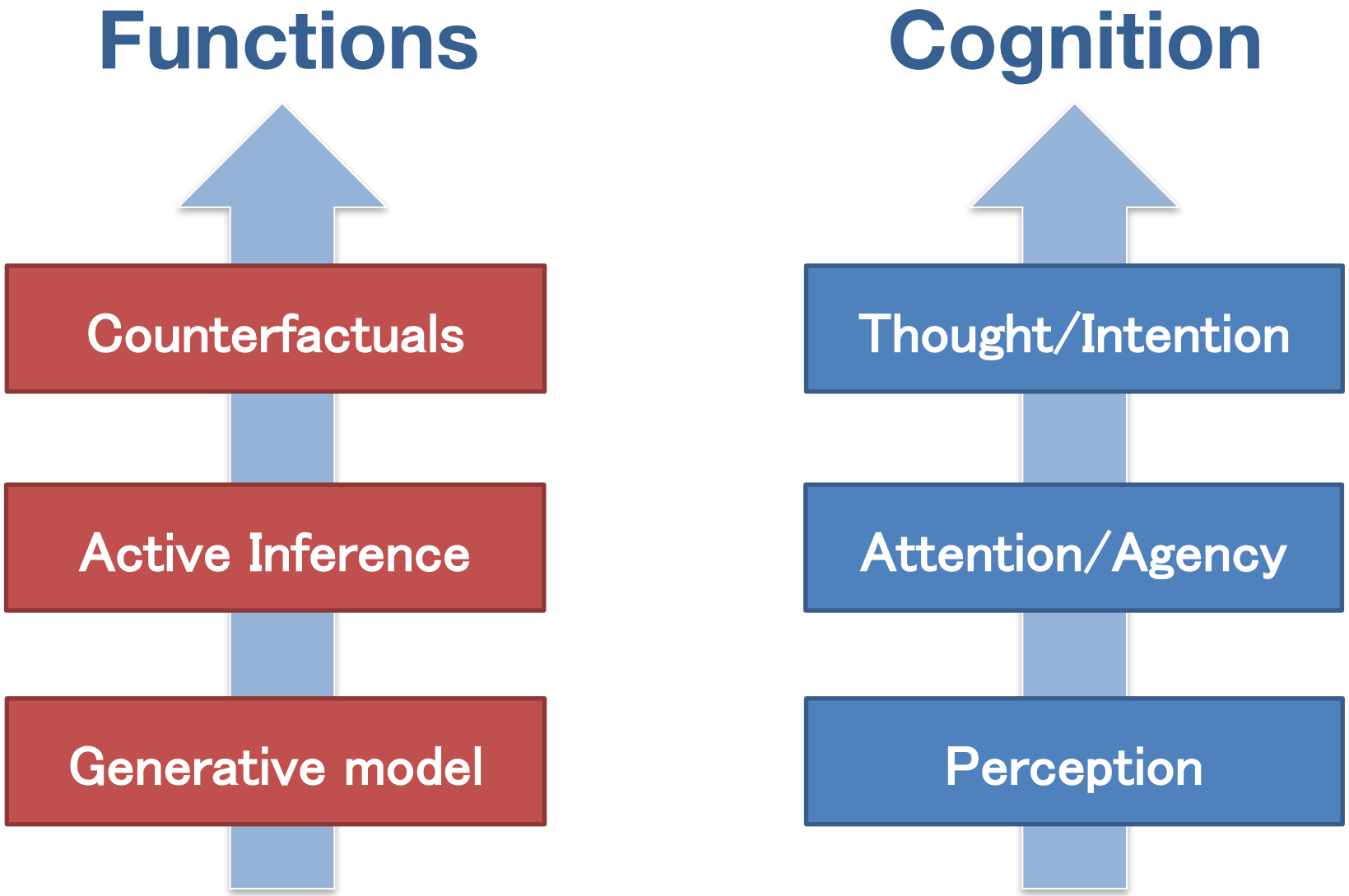
E.g. Aleksander (2005)

Scooped…

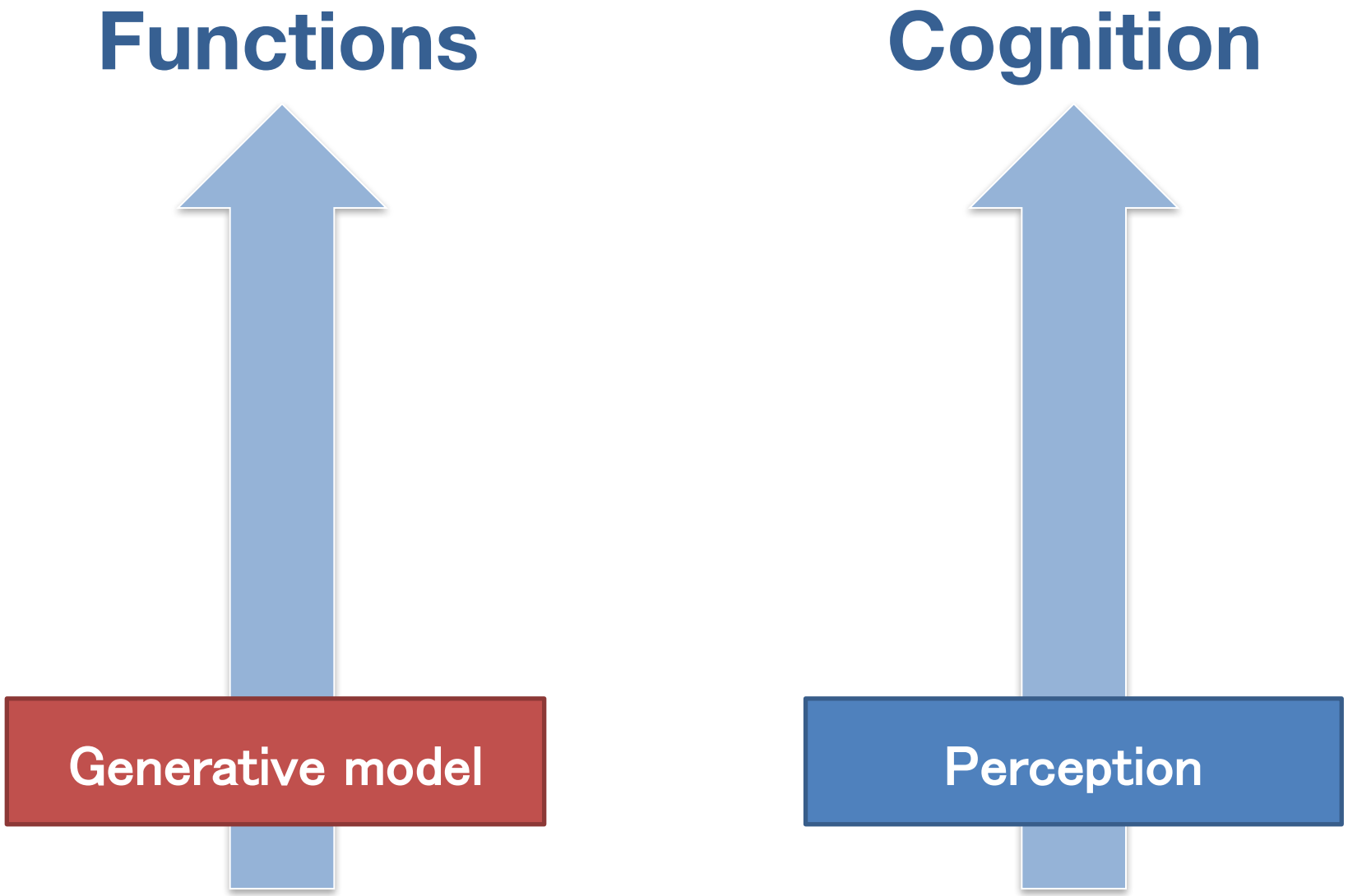
The evolution of the capacity to simulate seems to have culminated in subjective consciousness. . . . Perhaps consciousness arises when the brain's simulation of the world becomes so complete that it must include a model of itself – Dawkins (1976)

Implementing counterfactuals

Three key steps toward consciousness



Three key steps toward consciousness



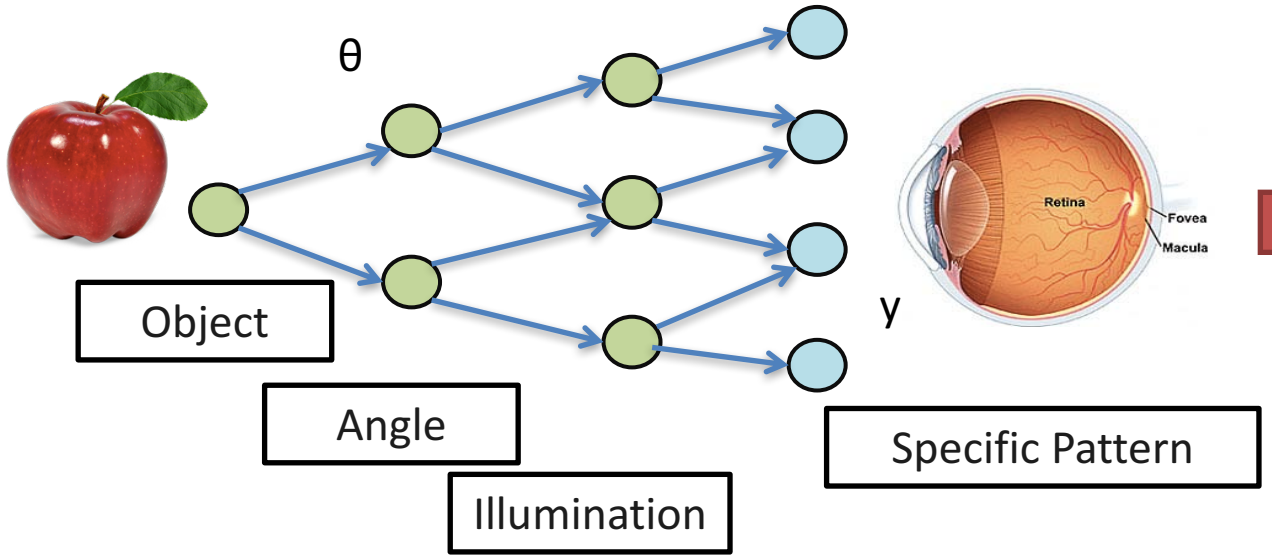
Learning generative models

Structure in the world

Sensory Input

Perception as Inference

$$F = -\ln p(\tilde{y}) + D(q(\vartheta;\lambda)\|p(\vartheta|\tilde{y}))$$



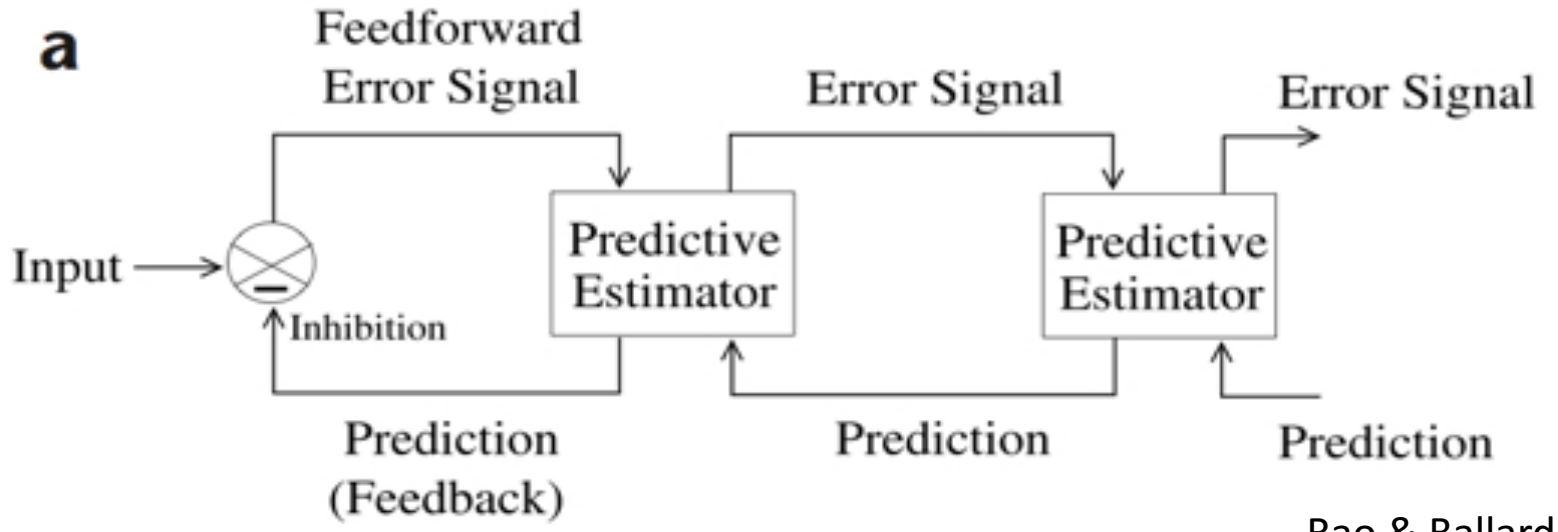
Functions achieved

Object recognition（Deep Learning）

Model of the brain itself (cf. Cleeremans)

Perception as Inference:
Predictive Coding (Free Energy Principle)

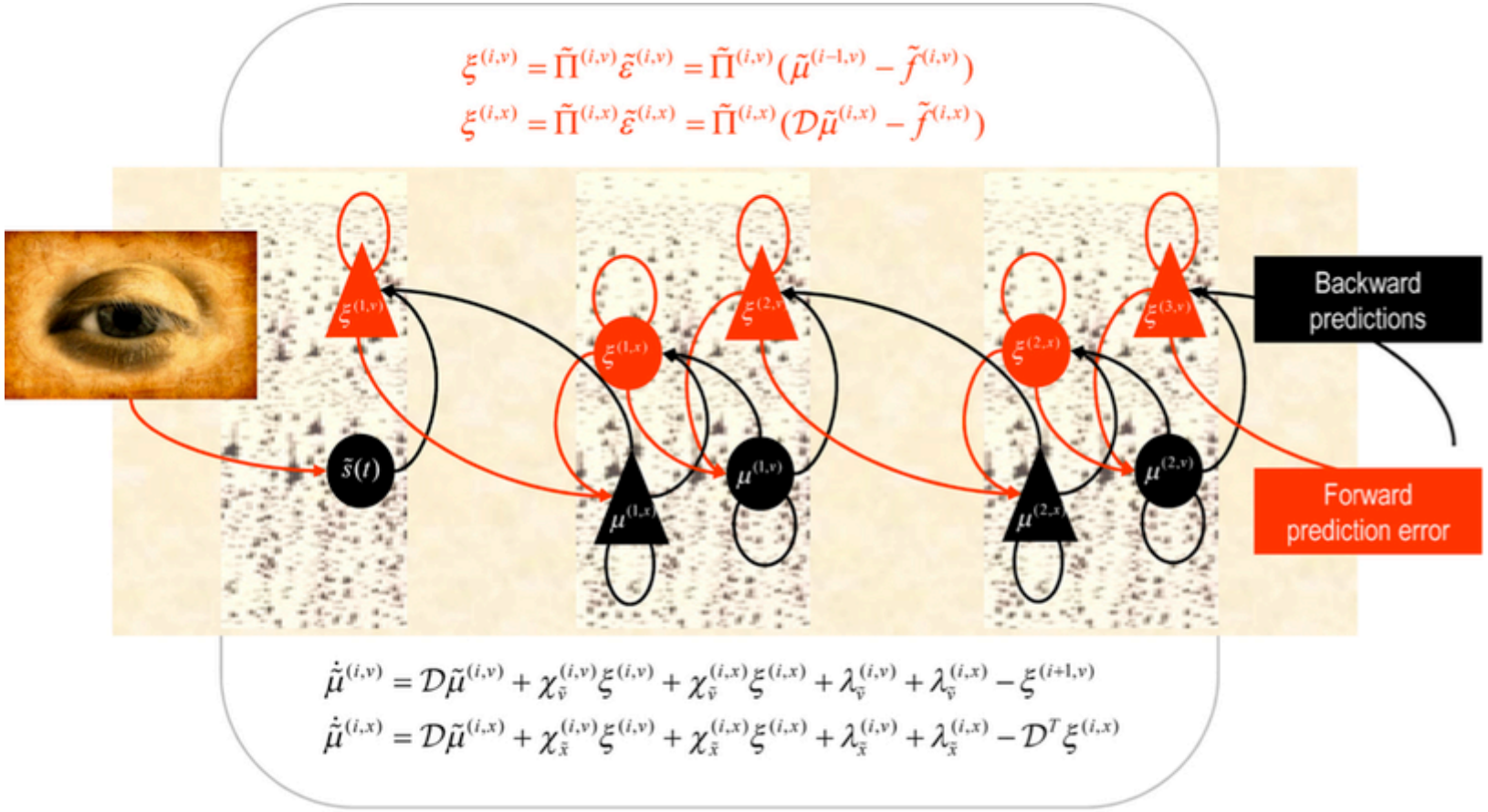
- A framework for perception as (Bayesian) inference of the causes of external stimuli.



Rao & Ballard (1999)

$$E_1 = \frac{1}{\sigma^2}(\mathbf{I} - f(\mathbf{U}\mathbf{r}))^{\mathrm{T}}(\mathbf{I} - f(\mathbf{U}\mathbf{r})) + \frac{1}{\sigma_{td}^2}(\mathbf{r} - \mathbf{r}^{td})^{\mathrm{T}}(\mathbf{r} - \mathbf{r}^{td})$$

Perception as Inference:
Predictive Coding (Free Energy Principle)



Hierarchical message passing in the brain

Learning statistical regularities of the environment

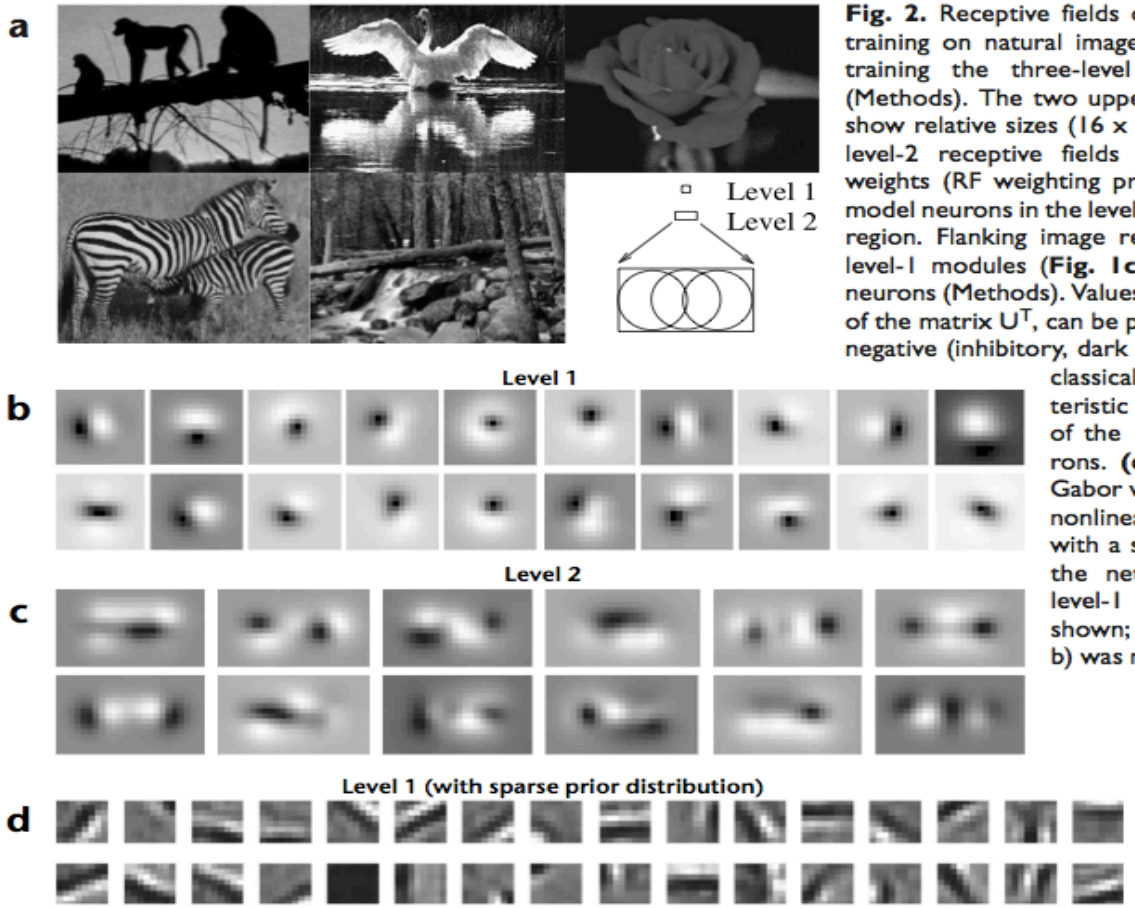


Fig. 2. Receptive fields of feedforward model neurons after training on natural images. **(a)** Five natural images used for training the three-level hierarchical network of Fig. 1c (Methods). The two upper boxes in the bottom right corner show relative sizes (16 x 16 and 16 x 26 pixels) of level-1 and level-2 receptive fields respectively. **(b)** Learned synaptic weights (RF weighting profiles) of 20 of the 32 feedforward model neurons in the level-1 module analyzing the central image region. Flanking image regions were analyzed by two other level-1 modules (**Fig. 1c**), each with 32 feedforward model neurons (Methods). Values for these synapses, which form rows of the matrix U^T, can be positive (excitatory, bright regions) or negative (inhibitory, dark regions). These RF profiles resemble classical oriented-edge/bar detectors characteristic of simple cells[2]. **(c)** RF profiles of 12 of the 128 level-2 feedforward model neurons. **(d)** Localized RF profiles resembling Gabor wavelets obtained by using a sigmoidal nonlinearity in the generative model, along with a sparse kurtotic prior distribution for the network activities (Methods). All 32 level-1 feedforward model neurons are shown; Gaussian windowing of inputs (as in b) was not necessary in this case.

RF profiles are learned internal models of the external environment.

Rao & Ballard (1999)

Going deeper with predictive coding

ARAYA
BRAIN IMAGING

Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning

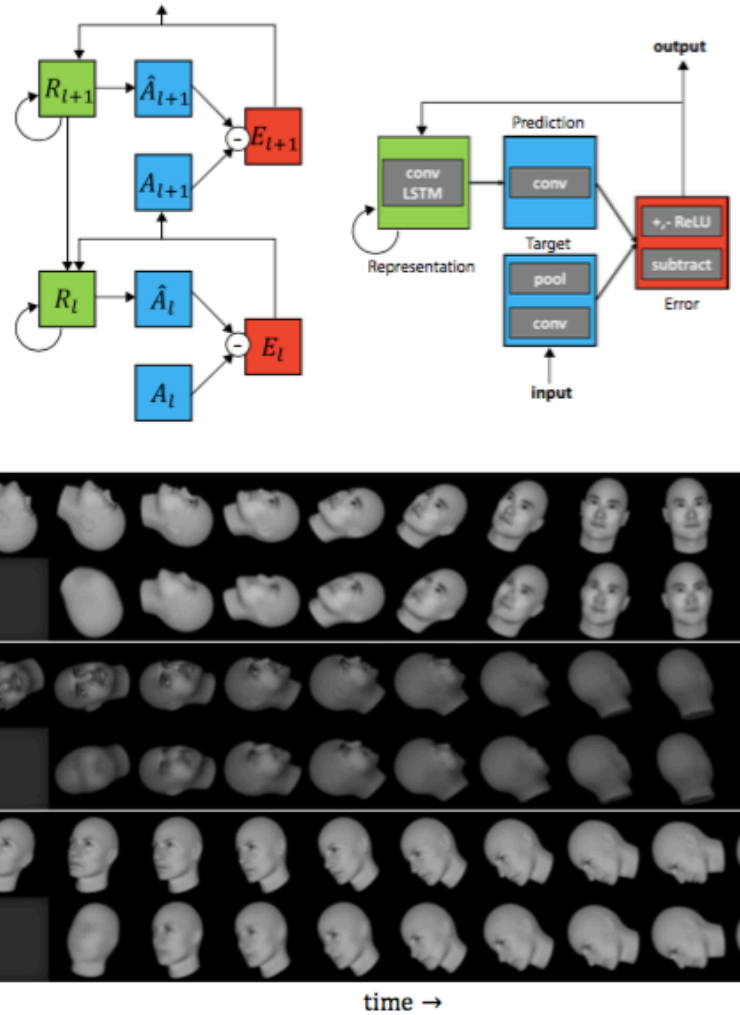
William Lotter
Harvard University
lotter@fas.harvard.edu

Gabriel Kreiman
Harvard University

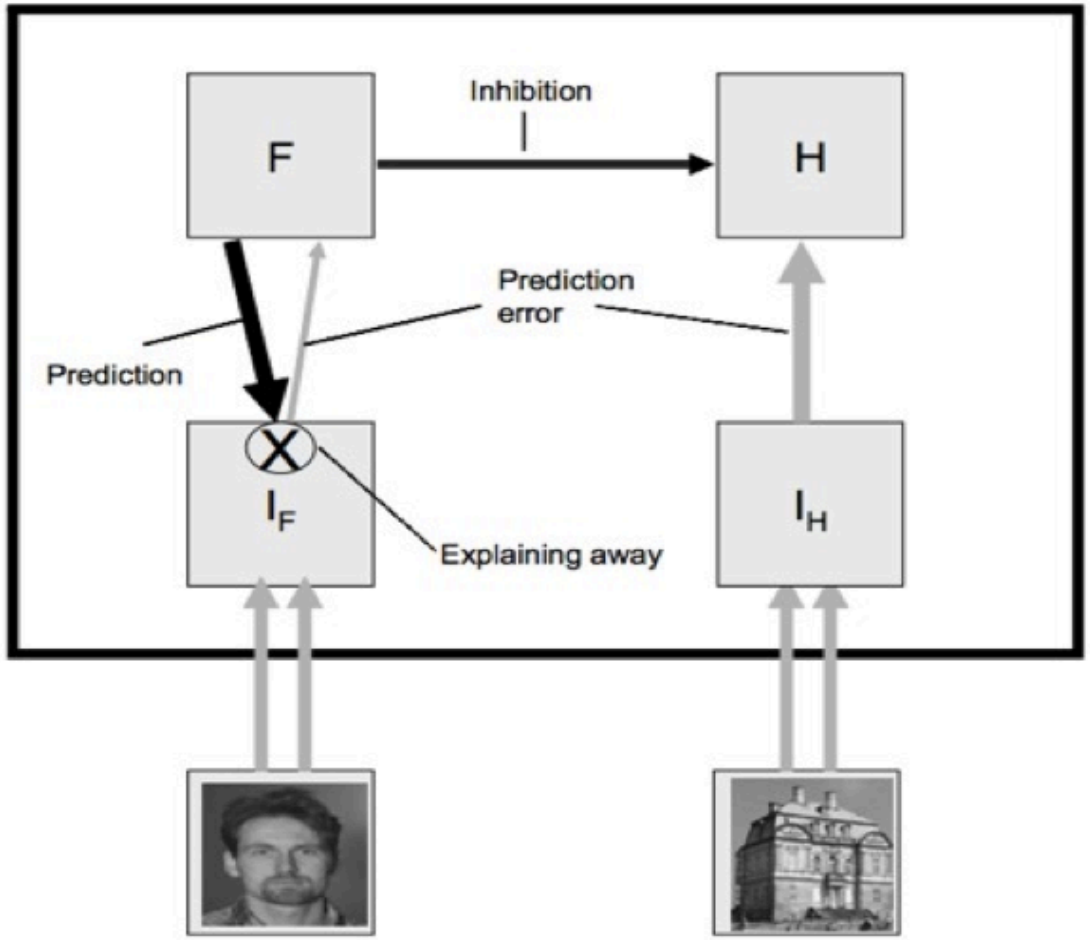
David Cox
Harvard University
davidcox@fas.harvard.edu

Abstract

While great strides have been made in using deep learning algorithms to solve supervised learning tasks, the problem of unsupervised learning — leveraging unlabeled examples to learn about the structure of a domain — remains a difficult unsolved challenge. Here, we explore prediction of future frames in a video sequence as an unsupervised learning rule for learning about the structure of the visual world. We describe a predictive neural network ("PredNet") architecture that is inspired by the concept of "predictive coding" from the neuroscience literature. These networks learn to predict future frames in a video sequence, with each layer in the network making local predictions and only forwarding deviations from those predictions to subsequent network layers. We show that these networks are able to robustly learn to predict the movement of synthetic (rendered) objects, and that in doing so, the networks learn internal representations that are useful for decoding latent object parameters (e.g. pose) that support object recognition with fewer training views. We also show that these networks can scale to complex natural image streams (car-mounted camera videos), capturing key aspects of both egocentric movement and the movement of objects in the visual scene, and generalizing across video datasets. These results suggest that prediction represents a powerful framework for unsupervised learning, allowing for implicit learning of object and scene structure. Accompanying code and video examples for the PredNet can be found at https://coxlab.github.io/prednet.

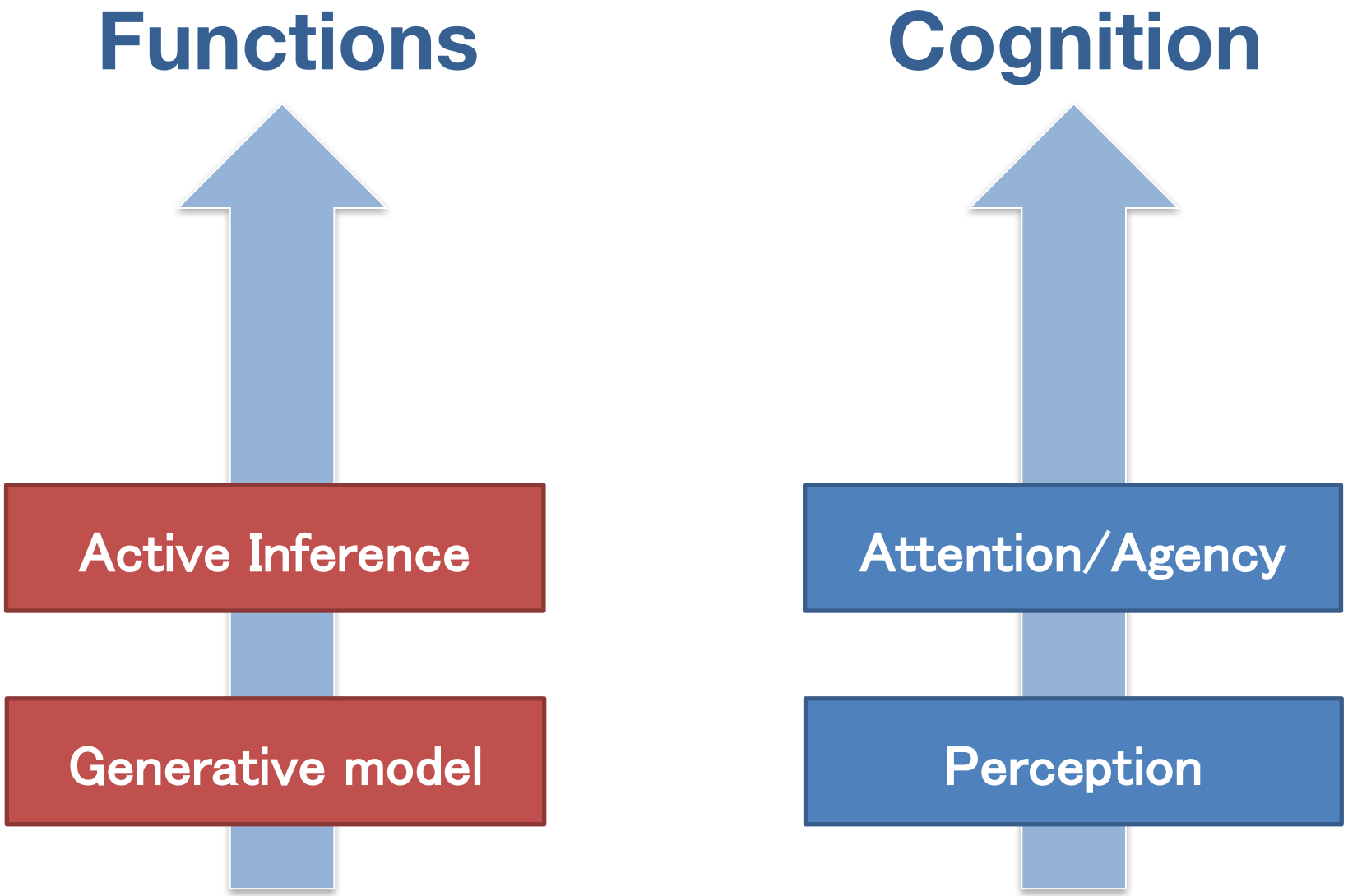


Predictive coding and rivalry



Hohwy, Roepstorff & Friston (2008)

Three key steps toward consciousness



Active inference

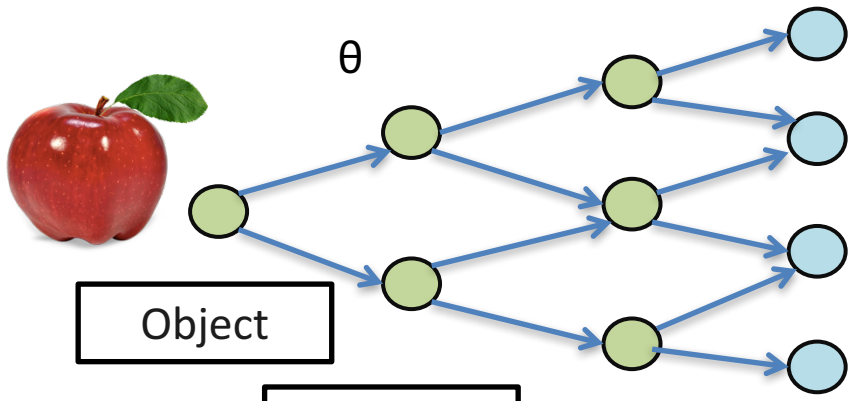
Prediction errors are minimized through action

Structure in the world

Sensory Input

Perception as Inference

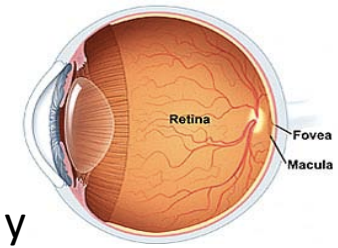
$$F = -\ln p(\tilde{y}) + D(q(\vartheta;\lambda)\|p(\vartheta|\tilde{y}))$$



Object

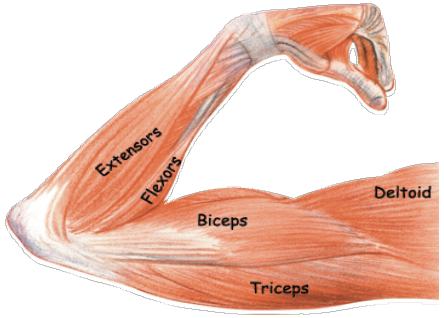
Angle

Illumination



y

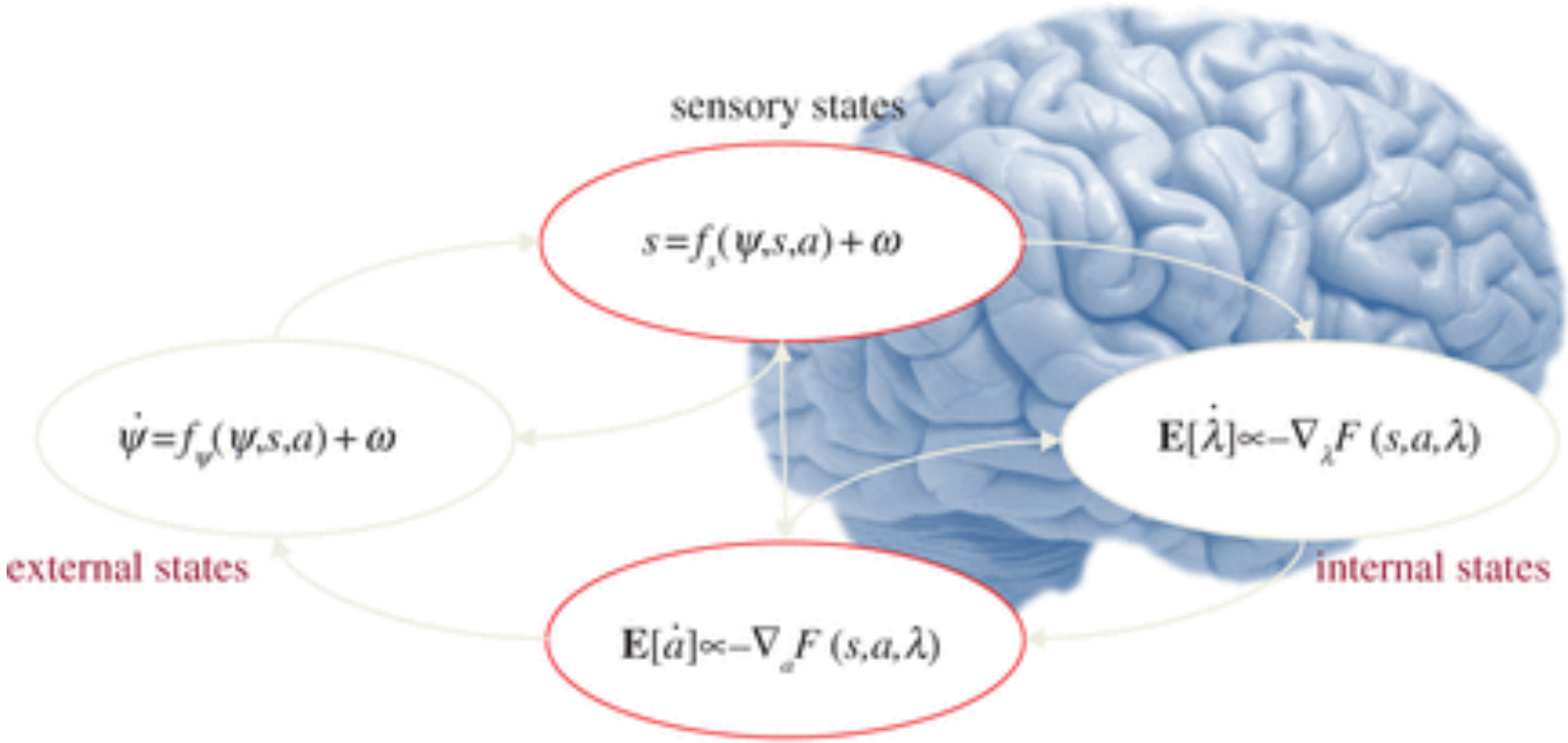
Specific Pattern



α

Action

Active inference



Active inference

ARAYA
BRAIN IMAGING

Functions achieved

Learning the structure of sensory motor contingency (Embodiment).

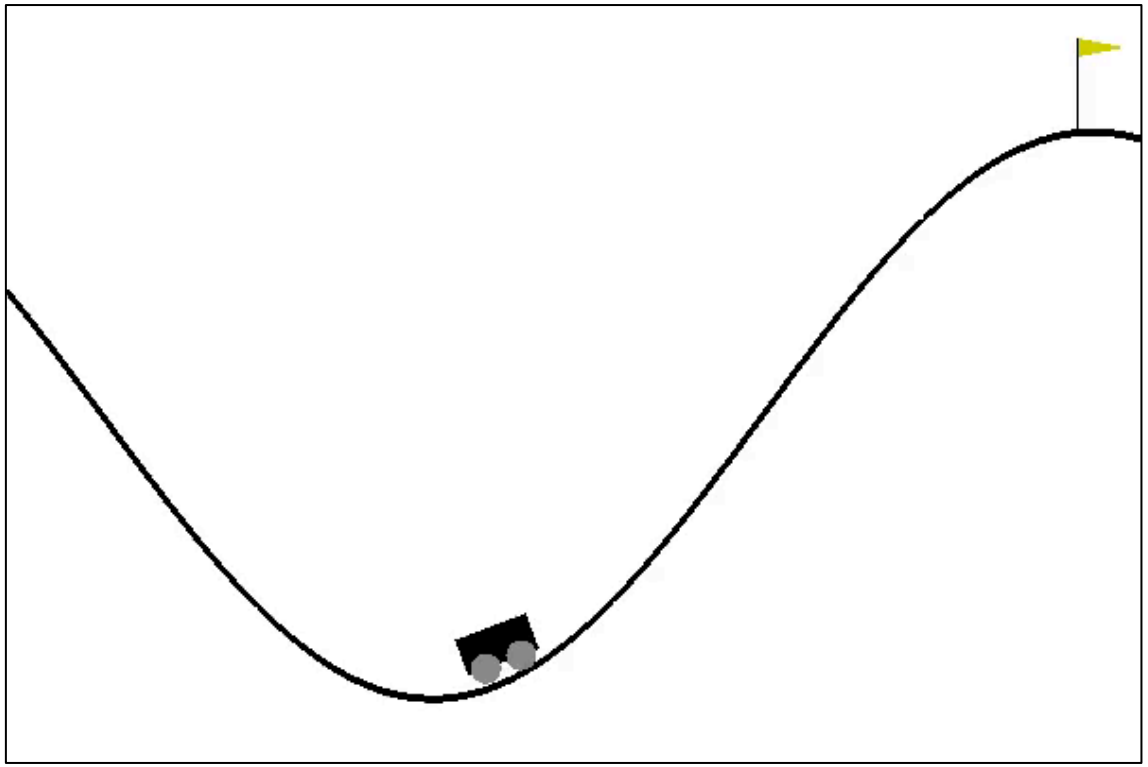
Learning the model of self through interaction with the environment.

Spontaneous behaviour from perceptual hypothesis.

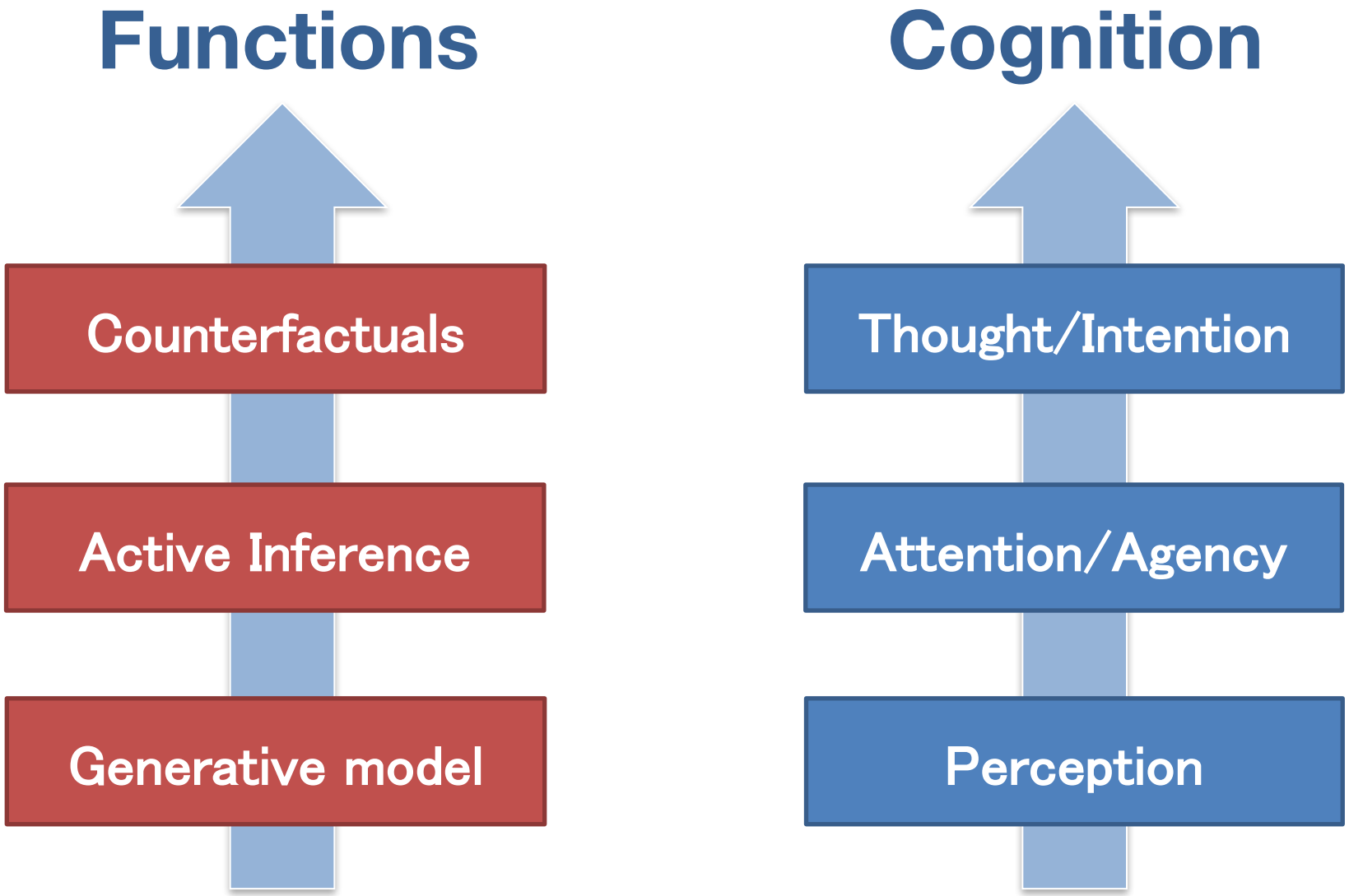
These are useful functions for building autonomous robots, cars, houses etc.

The Dark Room Problem

Active inference: agency in OpenAI Gym
(3) Idiosyncratic behaviour – suspension on slope



Three key steps toward consciousness



Infotaxis

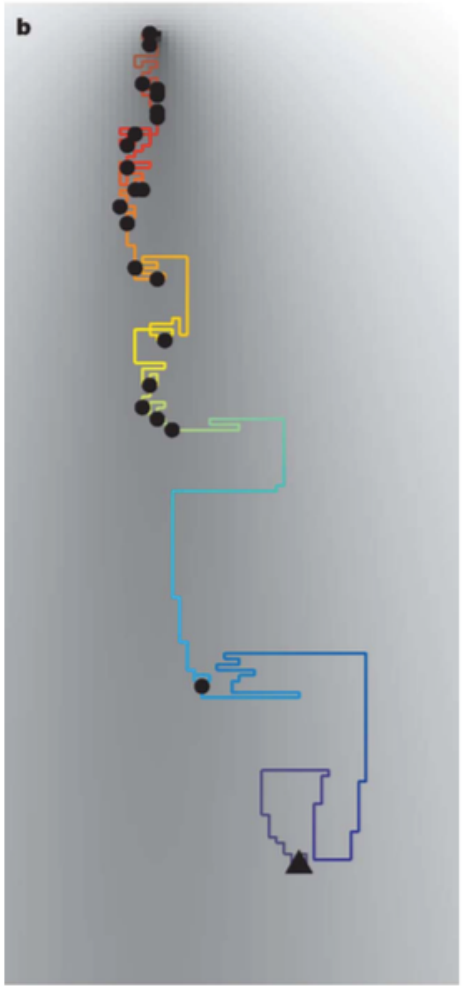
Action is chosen to minimize entropy

$$\overline{\Delta S}(\mathbf{r}\rightarrow\mathbf{r}_j) = P_t(\mathbf{r}_j)[-S] + [1-P_t(\mathbf{r}_j)][\rho_0(\mathbf{r}_j)\Delta S_0 + \rho_1(\mathbf{r}_j)\Delta S_1 + \ldots]$$

1st term: Probability of hitting the target.
2nd term: Otherwise

$P_k(r_j)$: probability of k detections in the next time bin Δt

Model based prediction of information gain.



Vergassola et al. (2007)

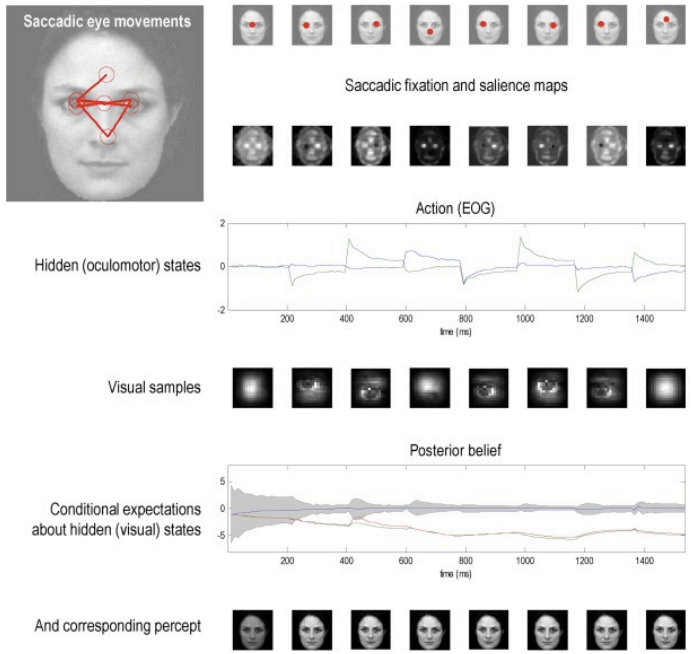
Saccade as experiments

ORIGINAL RESEARCH ARTICLE

Front. Psychol., 28 May 2012 | http://dx.doi.org/10.3389/fpsyg.2012.00151

Perceptions as hypotheses: saccades as experiments

Karl Friston[1]*, Rick A. Adams[1], Laurent Perrinet[1,2] and Michael Breakspear[3]



Next eye positions are chosen to test hypotheses generated by perception.

$$\tilde{\eta}_u(t) = \arg\max_{\tilde{\eta}_j} S(\tilde{\eta}_j)$$

$$S(\tilde{\eta}_j) = -H[q(\tilde{\psi} \mid \tilde{\mu}_x(t+\tau), \tilde{\mu}_v(t+\tau), \tilde{\eta}_j)]$$

$$= \tfrac{1}{2} \ln \left| \partial_{\tilde{\psi}} \tilde{\varepsilon}_j^T \Pi_\omega \partial_{\tilde{\psi}} \tilde{\varepsilon}_j + \Pi_\psi \right|$$

Counterfactuals everywhere

- Epistemic (intrinsic) value as opposed to extrinsic value.
- Exploration as opposed to exploitation.
- Model based reinforcement learning as opposed to model free reinforcement learning.
- Internal models (forward and inverse models) in motor control when disconnected from the plant.
- Infotaxis as opposed to chemotaxis.

Discussion point 3: Internal model

- What does it mean to have a model?
- Can we define internal models intrinsically?
- What kind of systems should be considered to possess an internal model?

Epistemic value

Expected free energy of a policy

$$(\hat{s}^*, \hat{\pi}^*) = \arg\min F(\tilde{o}, \hat{s}, \hat{\pi})$$

$$\Pr(a_t = u_t) = Q(u_t|\hat{\pi}^*)$$

$$F(\tilde{o}, \hat{s}, \hat{\pi}) = E_Q[-\ln P(\tilde{o}, \tilde{s}, \tilde{u}|m)] - H[Q(\tilde{s}, \tilde{u})]$$
$$= -\ln P(\tilde{o}|m) + D[Q(\tilde{s}, \tilde{u})||P(\tilde{s}, \tilde{u}|\tilde{o})]$$

$$\mathbf{Q}_\tau(\pi) = E_{Q(o_\tau, s_\tau|\pi)}[\ln P(o_\tau, s_\tau|\pi) - \ln Q(s_\tau|\pi)]$$
$$= E_{Q(o_\tau, s_\tau|\pi)}[\ln Q(s_\tau|o_\tau, \pi) + \ln P(o_\tau|m)$$
$$- \ln Q(s_\tau|\pi)] = \underbrace{E_{Q(o_\tau|\pi)}[\ln P(o_\tau|m)]}_{\text{Extrinsic value}}$$
$$+ \underbrace{E_{Q(o_\tau|\pi)}[D[Q(s_\tau|o_\tau, \pi)||Q(s_\tau|\pi)]]}_{\text{Epistemic value}}$$

Epistemic value is the expected information gain under predicted outcomes.

Friston et al. (2015)

Intention explained

- A functional interpretation of intention is the ability to form counterfactual predictions on the consequences of future actions.
- Intention is the ability to imagine what would happen.

Link to phenomenology of perceptual presence

- Perceptual presence as counterfactual predictions

- Phenomenological potentialities.

Seth (2014)

My reasons for the counterfactual theory

- Consciousness is tightly linked with non-reflexive behaviour.

- Explanation for intention

- The rationale behind the V1 hypothesis.

Intrinsic motivation requires counterfactuals

What is intrinsic motivation?

Conditions for Intrinsic Motivation

1. Independent of task
2. Computable from the subjective perspective of agent
3. Universal: Applicable to all kinds of sensorimotor loops.

(Salge, Glackin & Polani, 2014)

Empowerment

Behavioural Empowerment

- The adaptation brought about by natural evolution produced organisms that in absence of specific goals behave as if they were maximising their empowerment.

Evolutionary Empowerment

- The adaptation brought about by natural evolution produced organisms that in absence of specific goals behave as if they were maximising their empowerment.

AI Empowerment

- Empowerment provides a task-independent motivation that generate AI behaviour which is beneficial for a range of goal-oriented behaviour.

Salge, Glackin, & Polani (2013)

Empowerment

Definition:

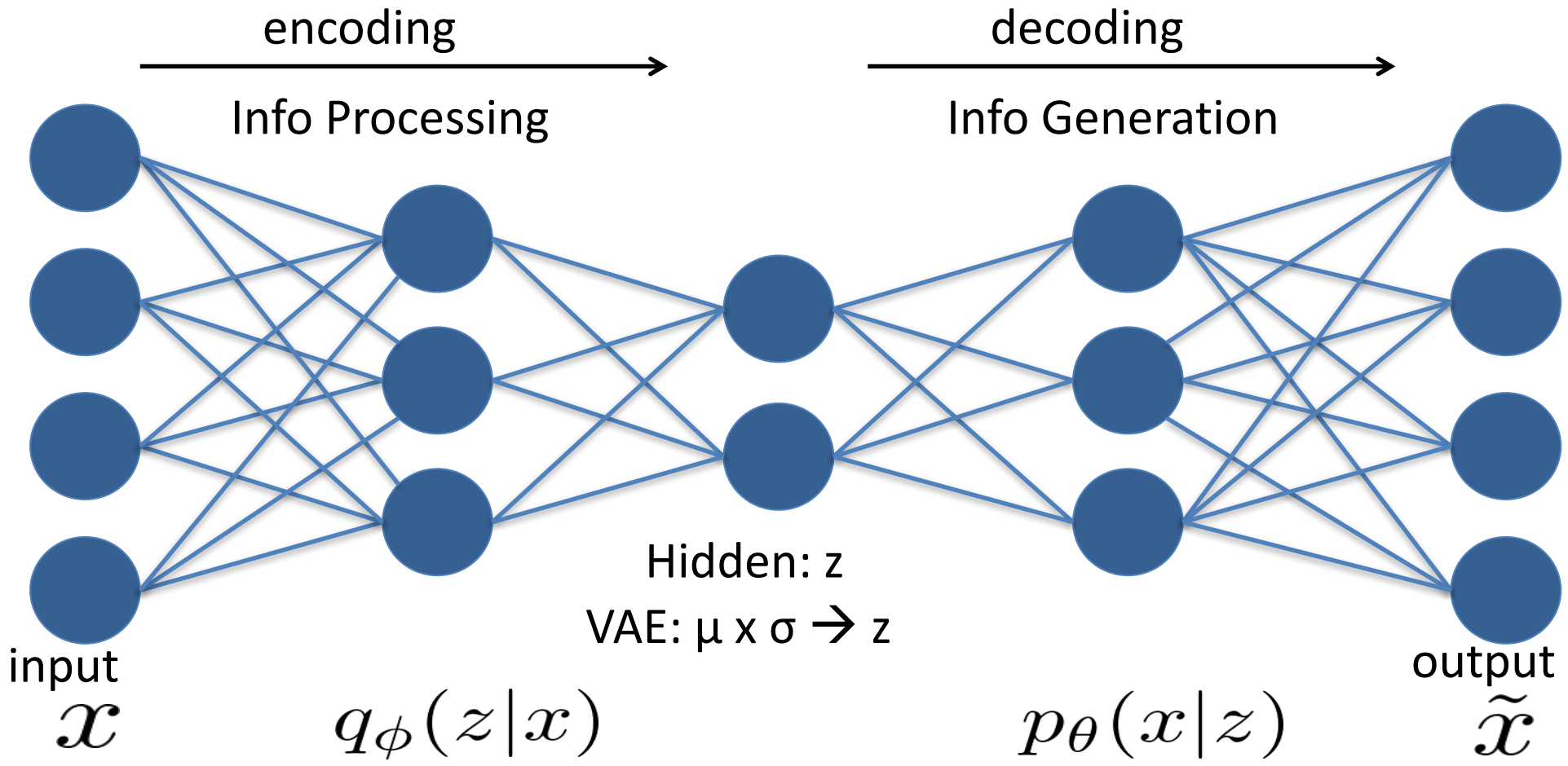
$$\mathfrak{E}_t = C\Big(p(s_{t+n}|a_t^n)\Big) = \max_{p(a_t^n)} I(A_t^n; S_{t+n}).$$

Empowerment quantifies "What the agent can potentially do".

Klyubin et al. (2005)

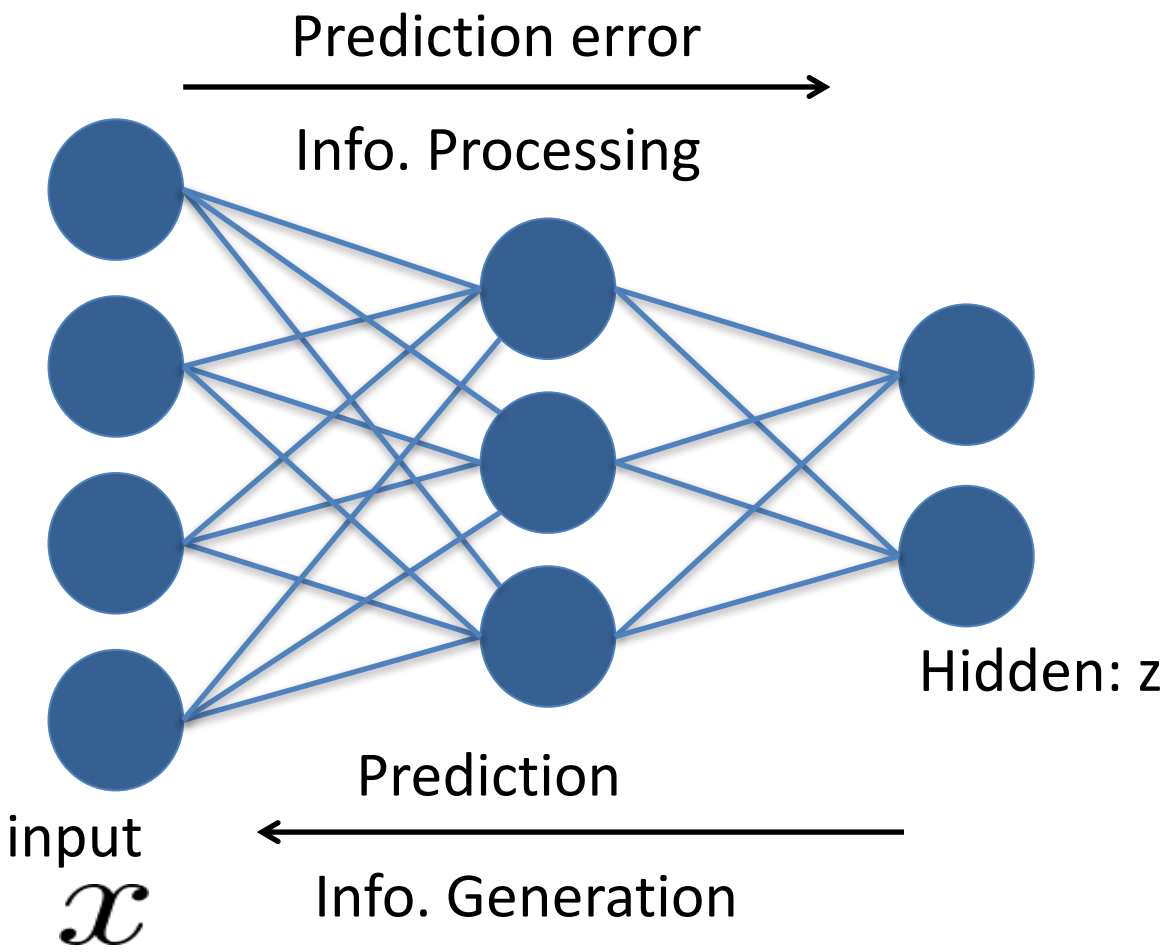
What does it mean to generate information?

Naïve Intuition



For individual layers, the analogy represented here is not obvious.

Predictive Coding

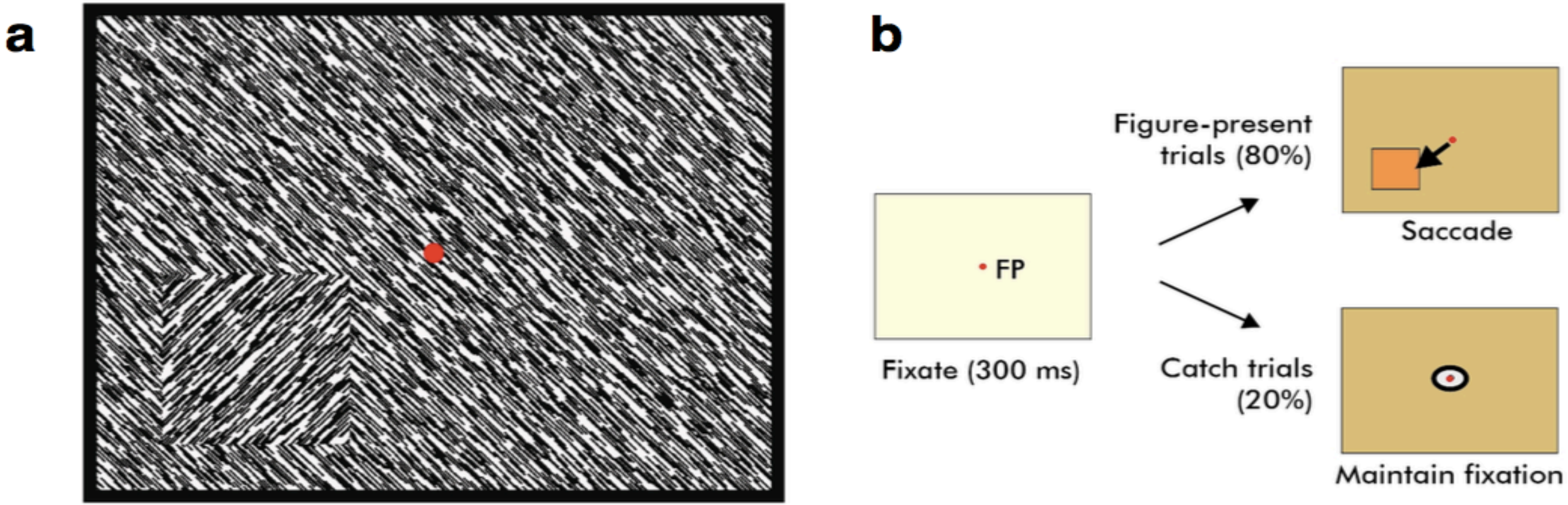


Encoding	**Decoding**
Processing	Generation
Compression	Decompression
Pred Err	Prediction
Feedforward	Feedback
Unconscious	Conscious

Empirical data support the notion that prediction corresponds to consciousness.

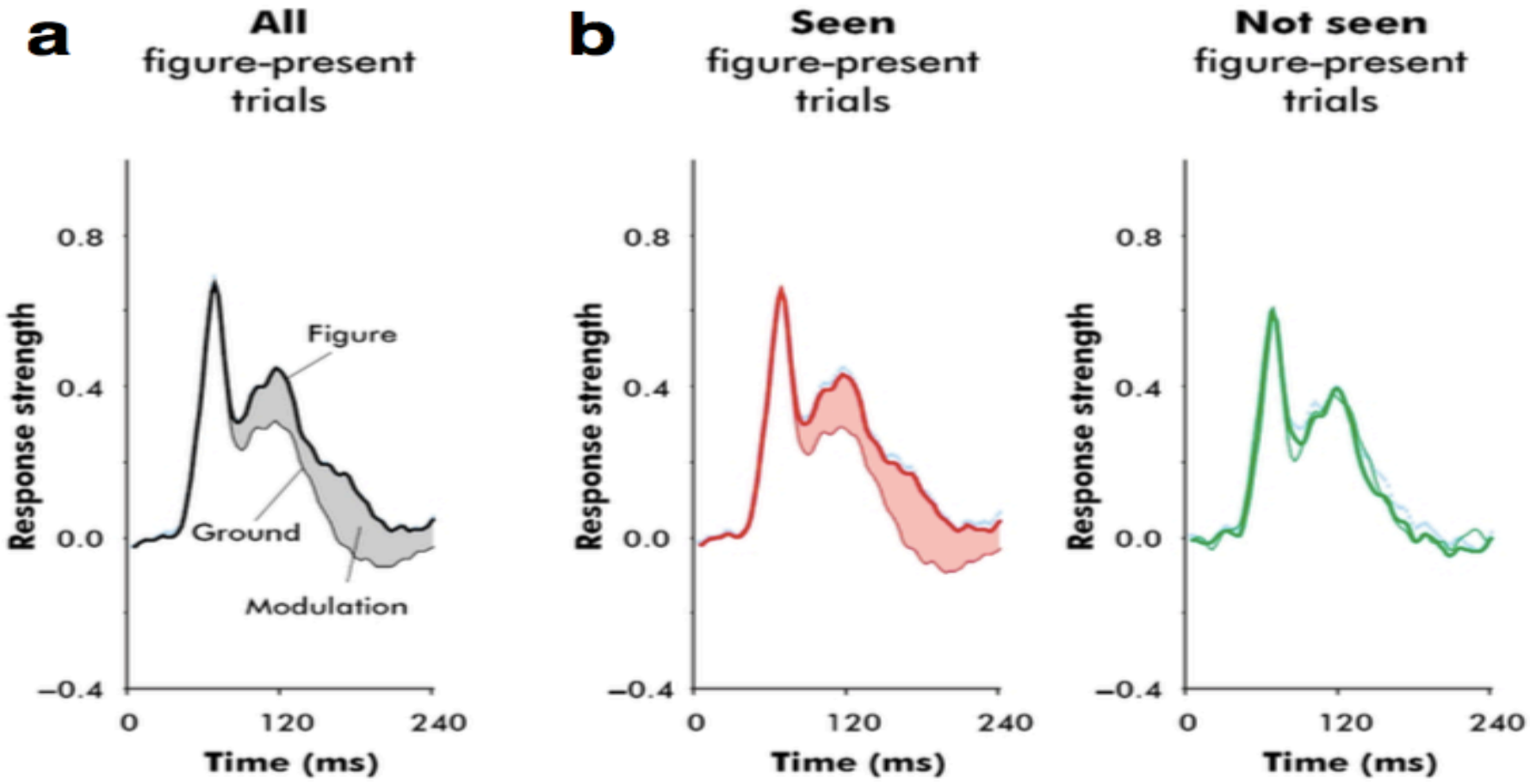
Feedback seems important for consciousness

Feedback and visual awareness



Super et al. (2001)

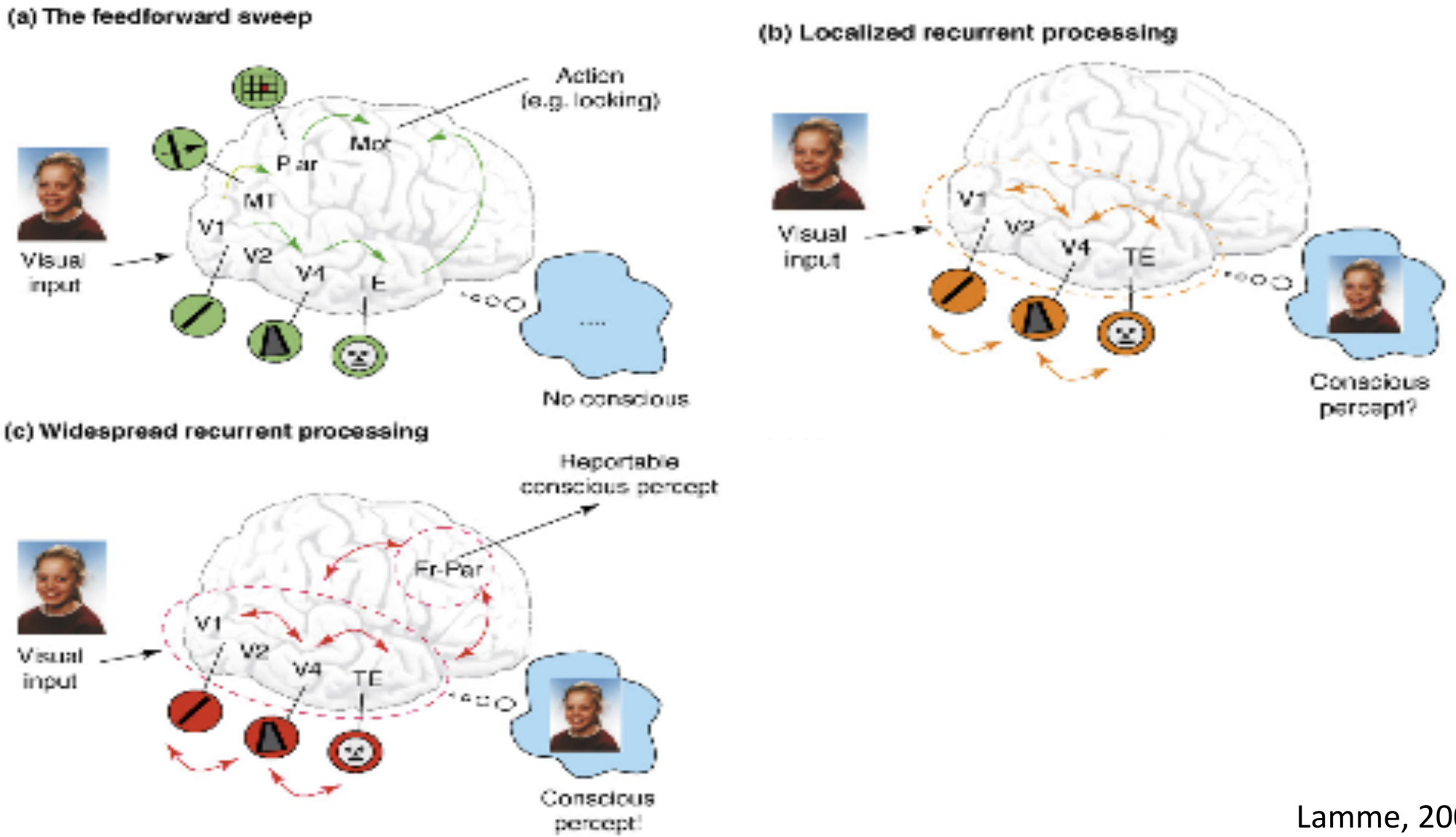
Feedback and visual awareness



見えたときだけ、遅いコンポーネントの活動が生じている。
これは、フィードバックによるものではないか？

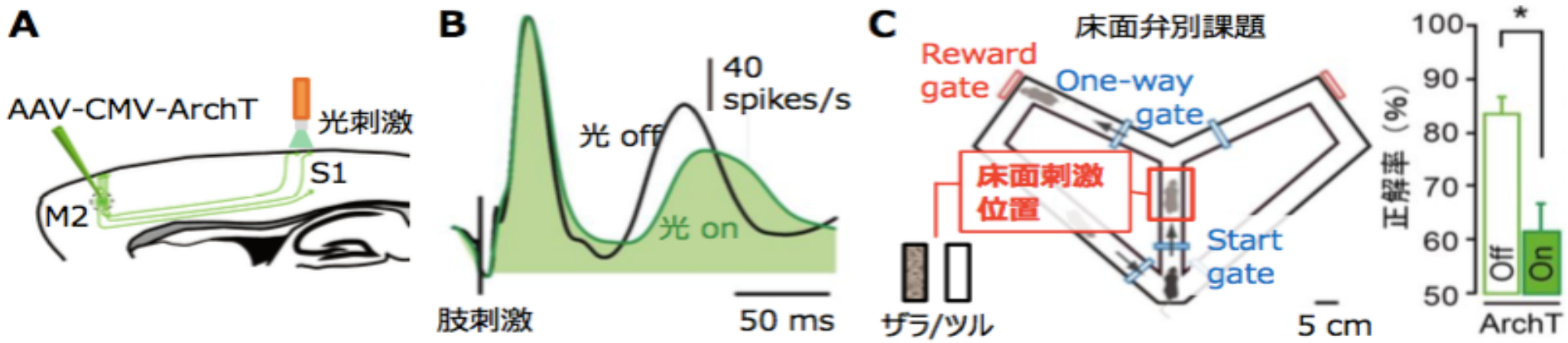
Super et al. (2001)

Feedback and visual awareness



Lamme, 2003

Optogenetics evidence

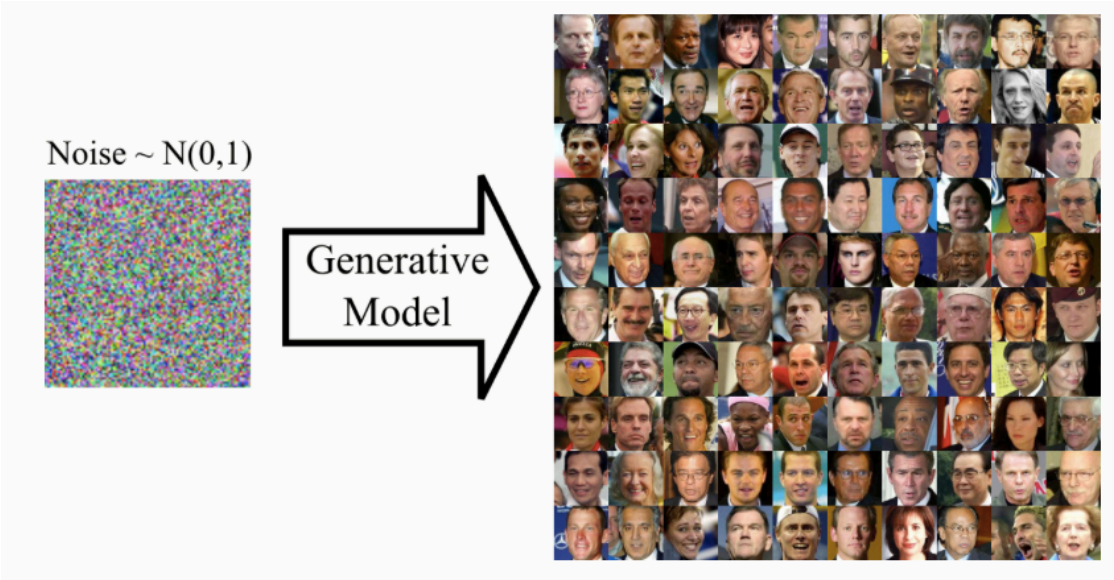


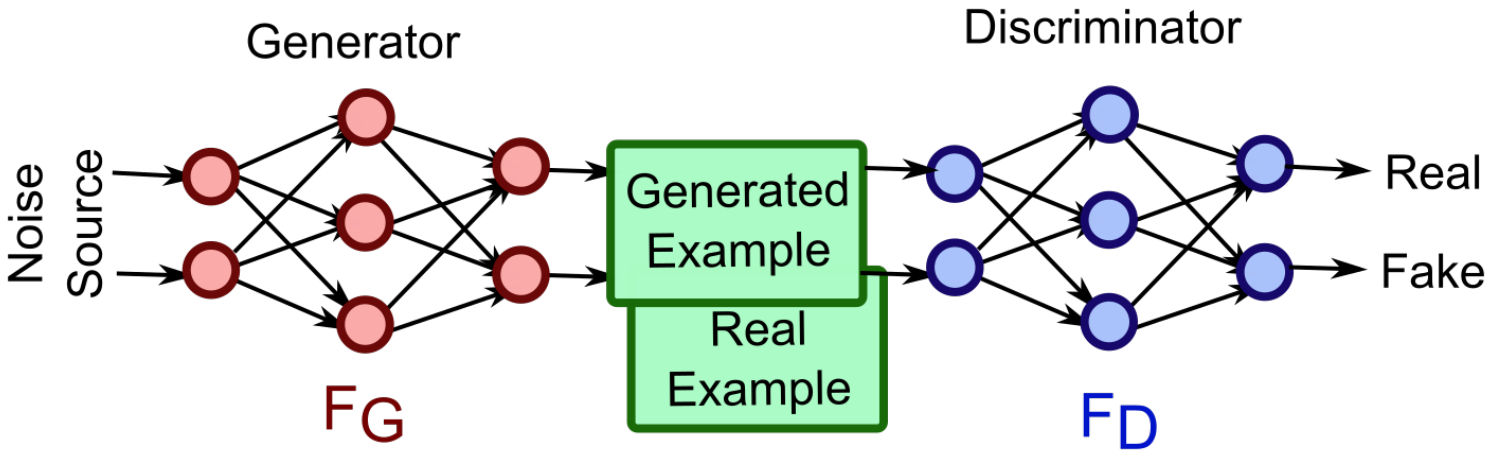
Suppression of M2-S1 feedback impairs tactile discriminability

Manita et al. (2015) Neuron

Creating an agent with information generation capability

Creating generative agents

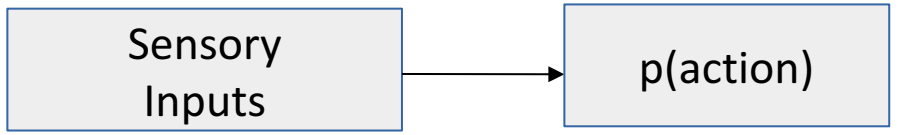




Learning control policy

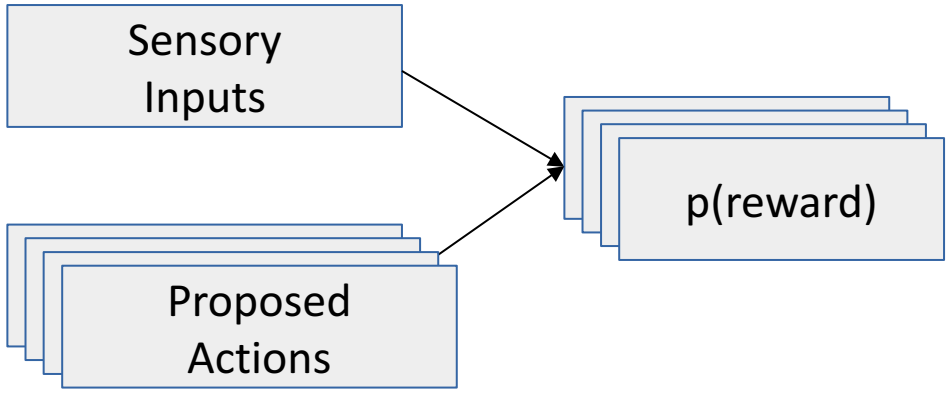
If there is a known example of correct behavior, can train a model to predict directly:

- 'What's the most likely action in this situation?'



Deep Q-Learning, Actor-Critic instead learn to predict:

- 'How much reward will I receive for this action?'
- Use to search for good actions.

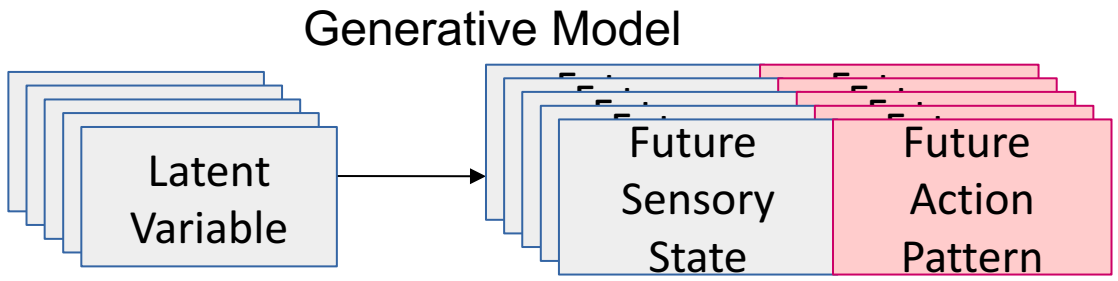


Generative control

Generative control:

- **'What is the distribution of possible futures?'**

Distribution includes actions → search for actions that lead to desirable futures



- Can change reward function on the fly
- Reward function can easily depend on things like uncertainties and surprisals
- Free long-term planning/coherency

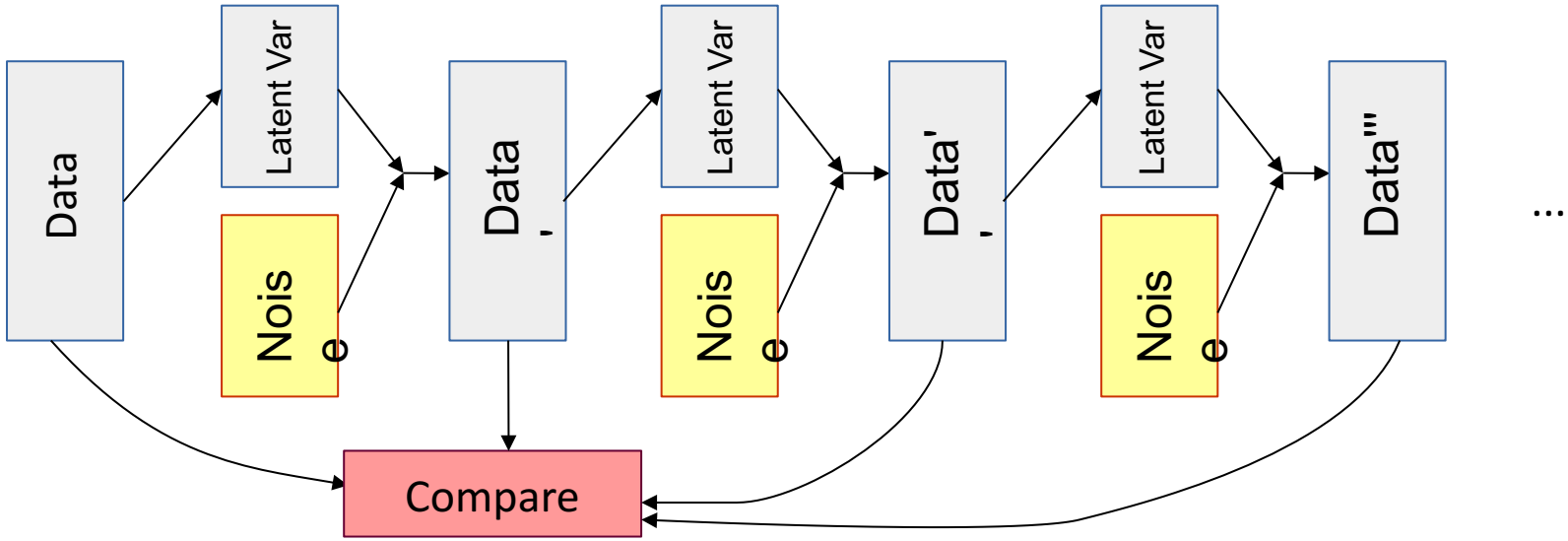
How to learn a generative model

- Parameter estimation given an explicit model (e.g. Bayesian Inference, Maximum Likelihood)
- Markov chains/Markov random fields

- Neural networks:
  - Generative Adversarial Networks (unstable!)
  - Variational Autoencoder ('blurry')
  - **Recurrent autoencoder**

Recurrent Autoencoder

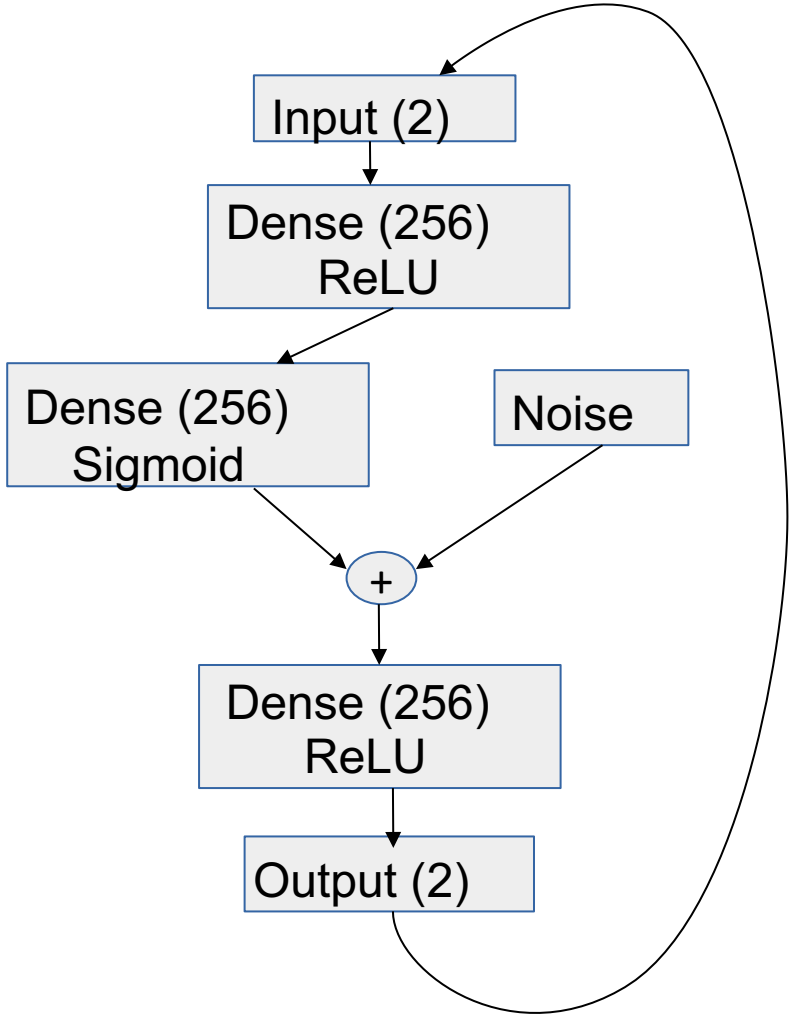
Repeated re-encoding allows an auto-encoder to better match the data distribution (Arulkumaran, Creswell, Bharath 2016)



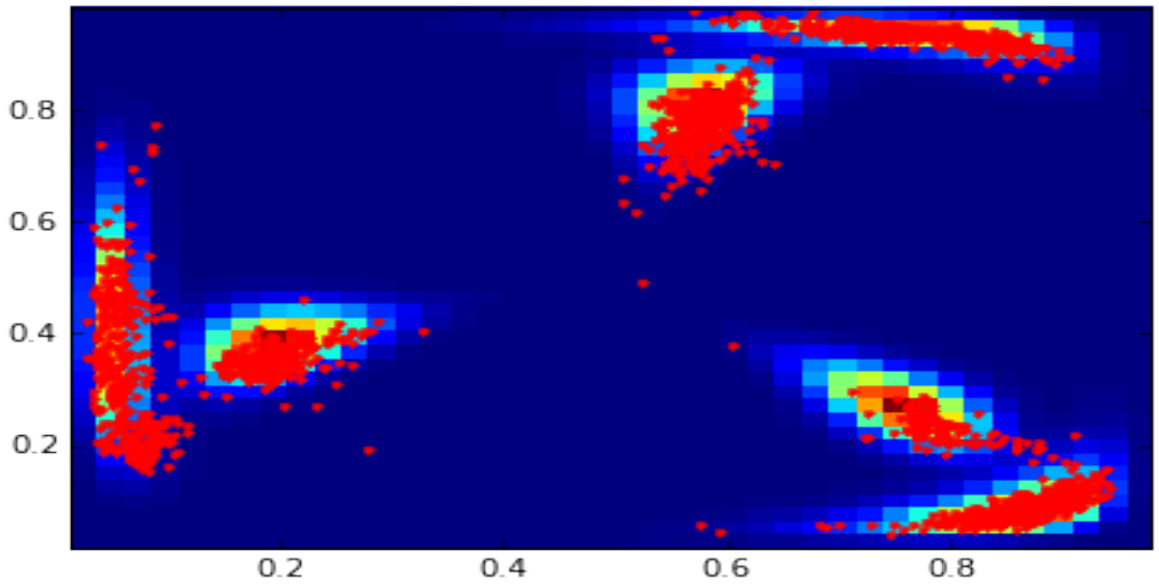
Generalize re-encoding as a recurrent neural network. Train while including the re-sampling process.

- Result learns to have fixed points corresponding to the data distribution

Test – Generate from Arbitrary 2D Distribution



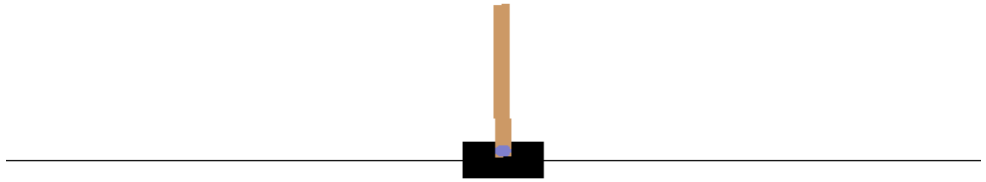
Checking whether this can learn a correct generative model

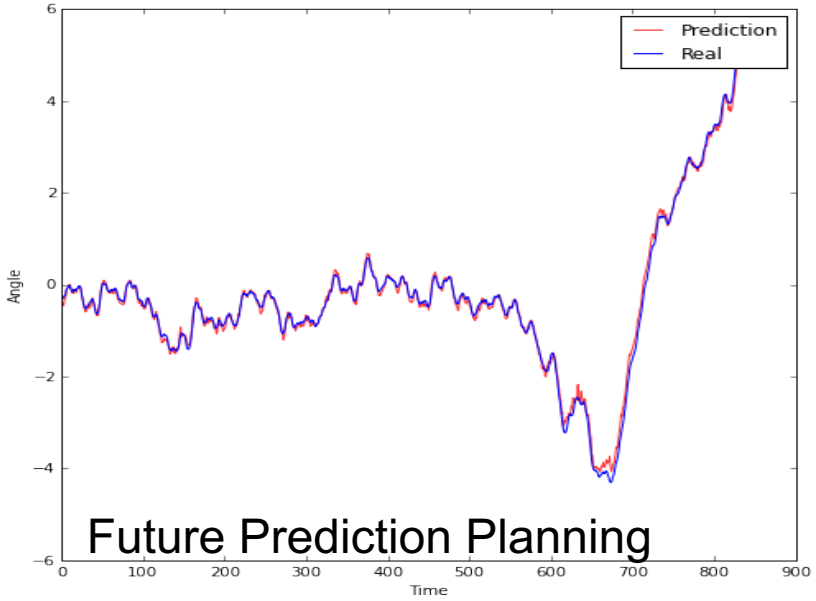


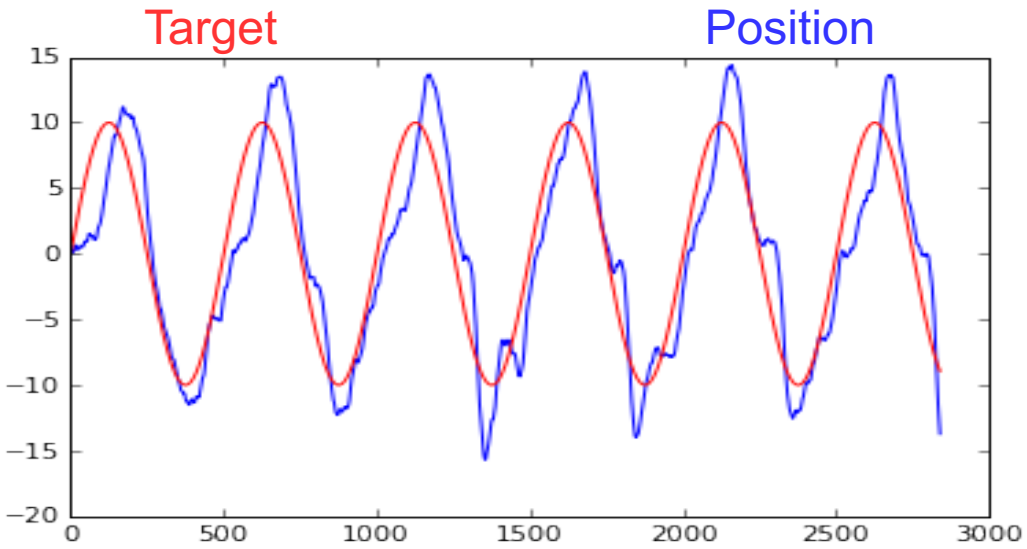
Generative Control of Cartpole

Learn action+sensor generative model, maps latent variable → prediction for next 16 frames

Gradient descent in latent space to pick action







Intrinsic Perspective

ARAYA BRAIN IMAGING

Shannon's Information Theory
(Extrinsic Perspective)

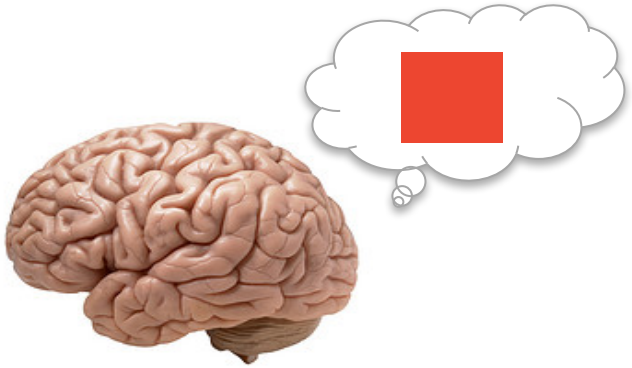
Stimulus

Neural Activity

Correspondences are assigned by the experimenter

Labels are given by an external observer

Integrated Information
Theory
(Intrinsic Perspective)



Brain can see only itself

No labels

Intrinsic information

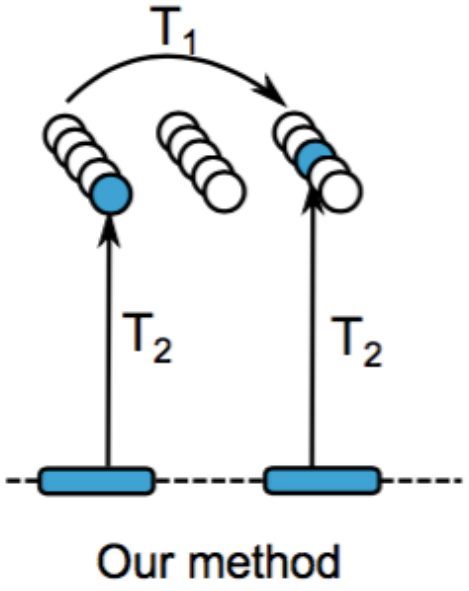
- Two meanings of intrinsicality
  - Self information
  - Physically grounded

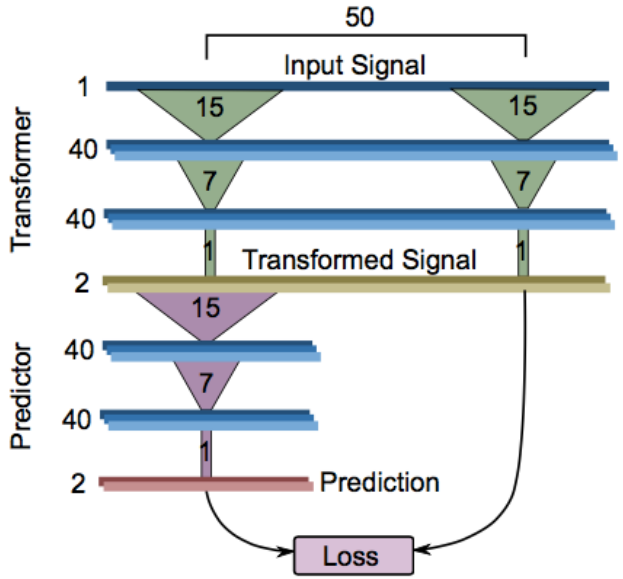
- What does it mean to compute intrinsically?
  - Epsilon machine (Crutchfield 2012)?
  - Neural Coarse Graining (Guttenberg, Biehl & Kanai 2016)?

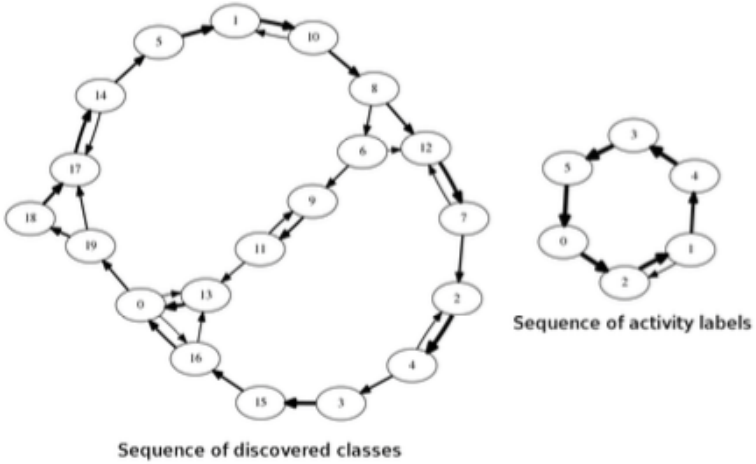
- What does it mean to have a model of the environment intrinsically?

Intrinsic computation

- Neural Coarse Graining







$$I_{pred} = H(X_{future}) - H(X_{future}|X_{past})$$

(Predictive Information; Non-trivial information closure)

Guttenberg, Biehl, & Kanai (2016)

Implicit & Explicit Counterfactuals

- Implicit Counterfactuals
  - View from a different angle (Perceptual Presence)
  - Massively parallel
  - Latent capacity to represent counterfactuals?

- Explicit Counterfactuals
  - Planning
  - Imagery
  - Episodic Memory
  - Sequential (Generative models need to be exploited)

Phenomenal vs Access

- Counterfactuals are opposite of qualia
  - Why is counterfactual important for phenomenal experience?

- Free floating qualia and self
  - Is the agency with counterfactual predictions a self?
  - Is an intrinsic internal model sufficient?

Future work

- GAN to generate future states.
- Non Bayesian NN CFP
  - Model space is too large
  - NN is good at constructing a model

- How does the acquisition of a generative model and CFP change phi in IIT?
  - Prediction is that phi suddenly increases.