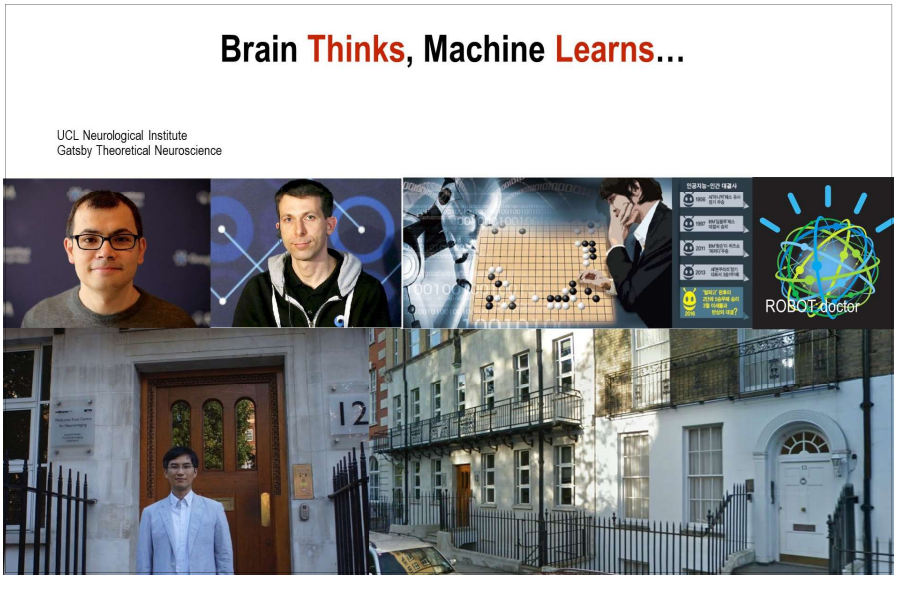
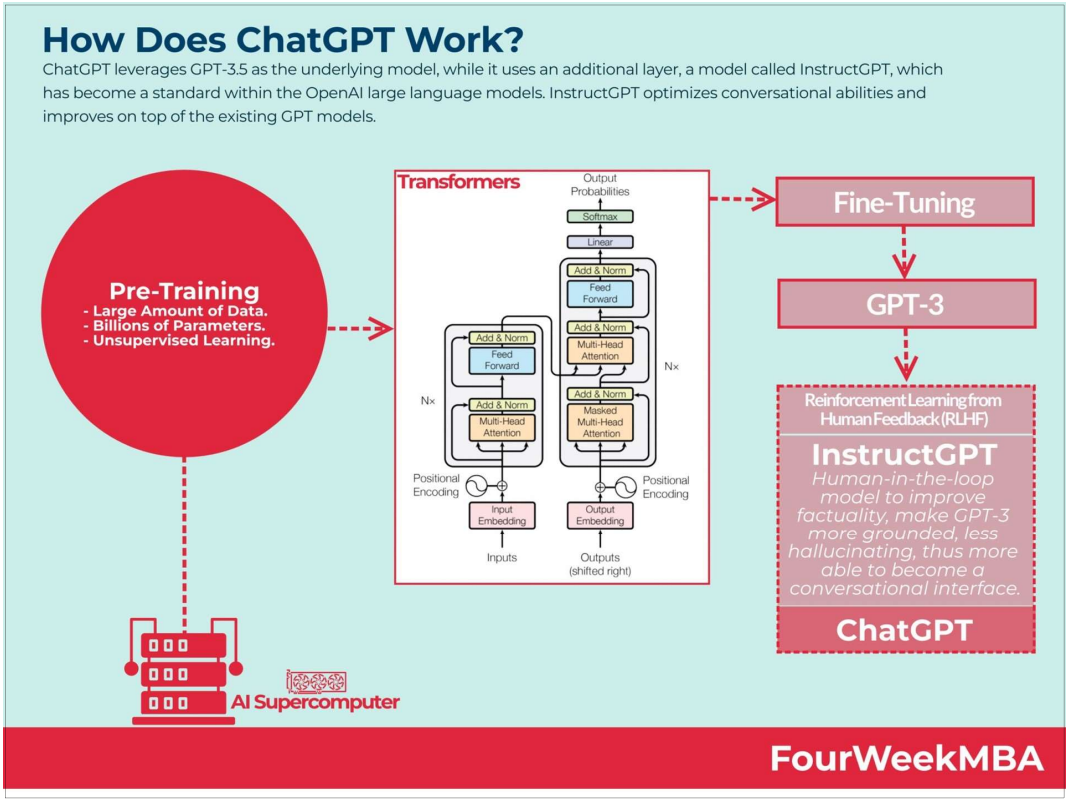
주제강연
02

ChatGPT의 뇌 인지 신경 과학적 함의

박해정 연세대학교 의학과 교수



"ChatGPT에 대해서 무엇을 물어야 하는가?"



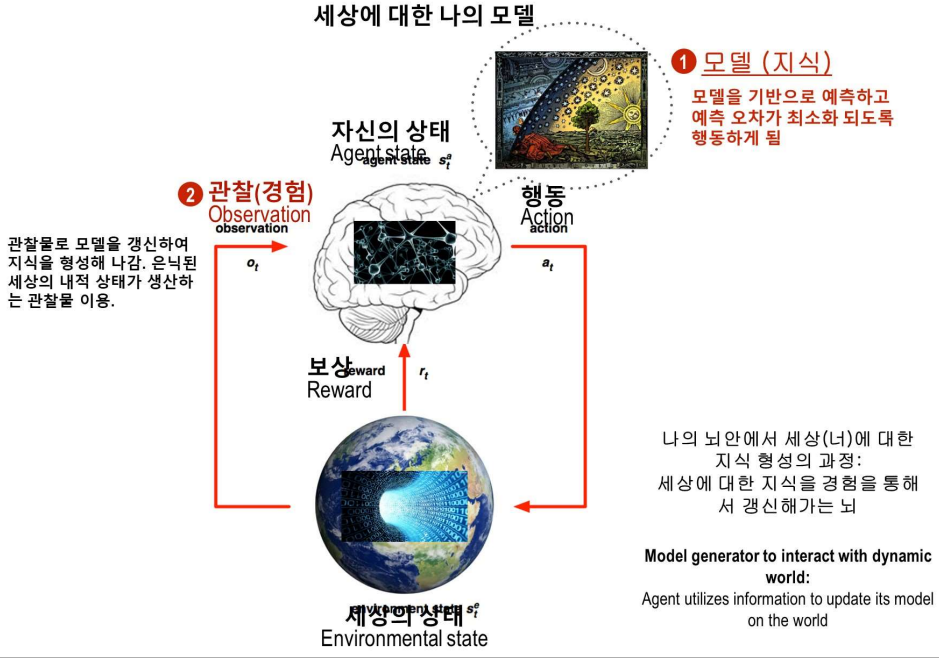
Generative Pre-trained Transformer

FREE ENERGY

LIKELIHOOD

뇌(腦)가 당면한 세상

변화하는 환경, 환경을 이해하는 뇌



나의 뇌안에서 세상(너)에 대한
지식 형성의 과정:
세상에 대한 지식을 경험을 통해
서 갱신해가는 뇌

Model generator to interact with dynamic world:
Agent utilizes information to update its model on the world

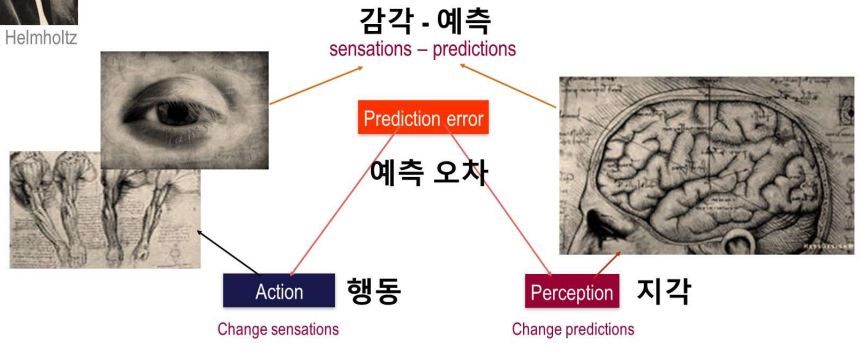
Hermann von Helmholtz

"Objects are always imagined as being present in the field of vision as would have to be there in order to produce the same impression on the nervous mechanism" - von Helmholtz

인식과 행동의 순환 과정 Perception-Action Cycle

Free energy principle as a general predictive coding

예측 부호화에 의한 자유에너지 원리



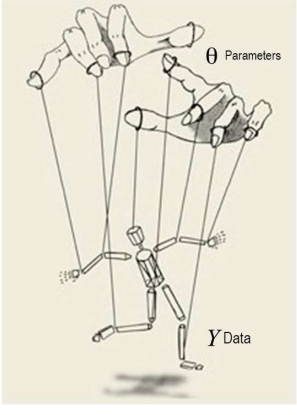
감각-예측 오차를 줄이기 위해 행동한다. 말이 되게 하기. 설명해 내기.

...prediction errors drive action and perception to suppress themselves

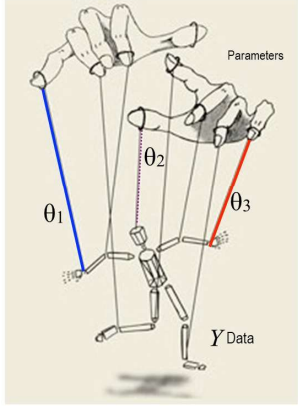
Prediction based on the generative model; according to sensory evidence, adjust prediction or adjust sensory evidence by actions to minimize **prediction errors** (surprise, free energy)

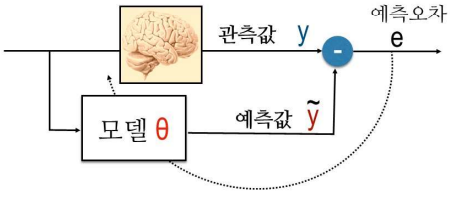
from Prof. Friston, K

생성모델

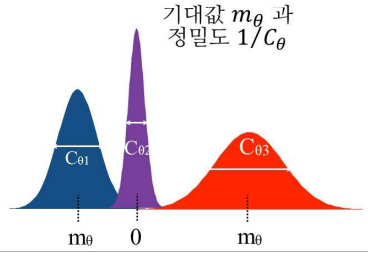


Prior : 매개변수 검색 범위와 강도를 선험적 한정





업데이트 (예측오차를 제거하는 방향으로)



베이지안 추론

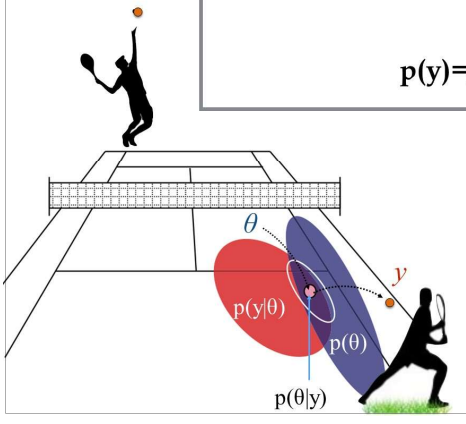
사전 지식을 바탕으로 관측물 y를 이용하여 은닉된 실체 θ 를 확률적으로 추론해 나감

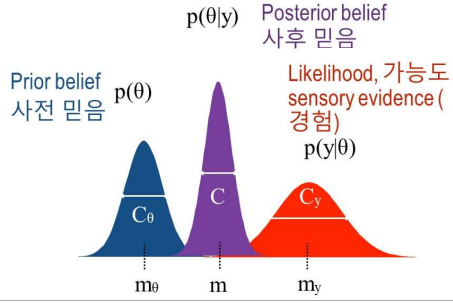
추론 **Inference (e.g. perception)** : p(θ|y) from observation y to hidden parameters θ

학습 **Learning (& memory)**: from a prior belief p(θ) to a posterior belief p(θ|y) using observation y

posterior $p(\theta \mid y) = \frac{p(y \mid \theta)\, p(\theta)}{p(y)}$ likelihood, model prior, evidence

$p(y) = \int p(y \mid \theta')\, p(\theta')\, d\theta'$ marginalisation





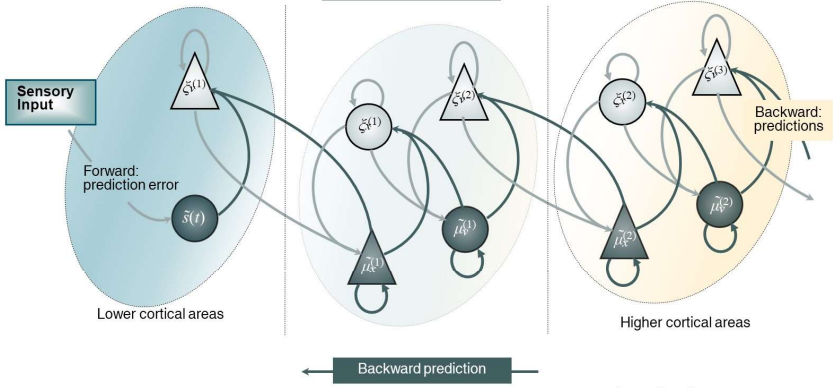
What and how information is delivered over the rich-club like network?

Predictive coding

$$\xi_v^{(i)} = \Pi_v^{(i)} \varepsilon_v^{(i)} = \Pi_v^{(i)} \left(\mu_v^{(i-1)} - g\left(\mu^{(i)}\right) \right)$$

$$\xi_x^{(i)} = \Pi_x^{(i)} \varepsilon_x^{(i)} = \Pi_x^{(i)} \left(D\mu_x^{(i)} - f\left(\mu^{(i)}\right) \right)$$

Forward prediction error



Backward prediction

Synaptic plasticity

$$\ddot{\mu}_{\theta_{ij}} = -\partial_{\theta_{ij}} \varepsilon^T \xi$$

$$\dot{\mu}_v^{(i)} = D\mu_v^{(i)} - \left(\partial_v \varepsilon^{(i)}\right)^T \xi^{(i)} - \xi_v^{(i+1)}$$

$$\dot{\mu}_x^{(i)} = D\mu_x^{(i)} - \left(\partial_x \varepsilon^{(i)}\right)^T \xi^{(i)}$$

Synaptic gain

$$\ddot{\mu}_{\gamma_i} = \frac{1}{2} tr\left(\partial_{\gamma_i} \Pi \left(\xi \xi^T - \Pi\left(\mu_\gamma\right) \right) \right)$$

Generative Pre-trained Transformer

NETWORK TOPOLOY

MODEL, PRIOR

Reducing the Dimensionality of Data with Neural Networks

G. E. Hinton* and R. R. Salakhutdinov

High-dimensional data can be converted to low-dimensional codes by training a multilayer neural network with a small central layer to reconstruct high-dimensional input vectors. Gradient descent can be used for fine-tuning the weights in such "autoencoder" networks, but this works well only if the initial weights are close to a good solution. We describe an effective way of initializing the weights that allows deep autoencoder networks to learn low-dimensional codes that work much better than principal components analysis as a tool to reduce the dimensionality of data.

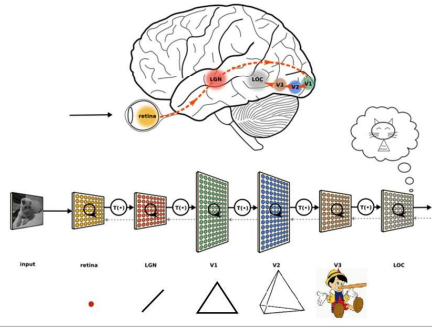
Dimensionality reduction facilitates the classification, visualization, communication, and storage of high-dimensional data. A simple and widely used method is principal components analysis (PCA), which finds the directions of greatest variance in the data set and represents each data point by its coordinates along each of these directions. We describe a nonlinear generalization of PCA that uses an adaptive, multilayer "encoder" network

Science 313(2006):504-507

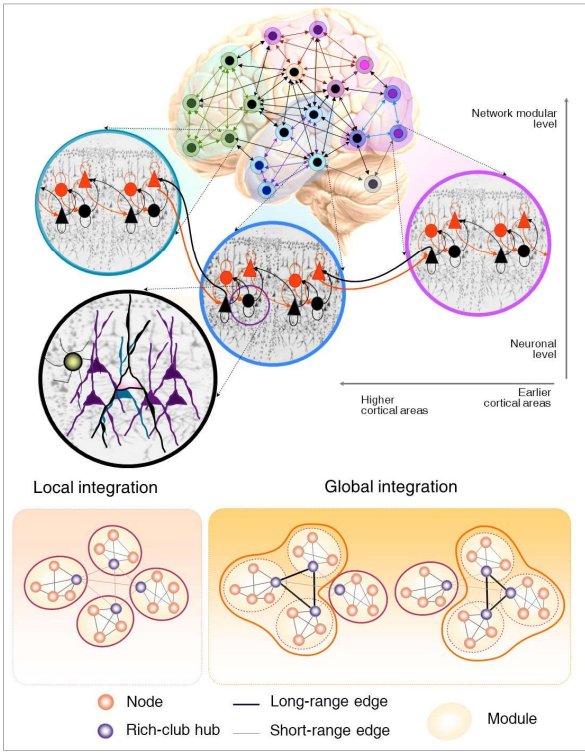
심층 신경망 (Deep Neural Nets)

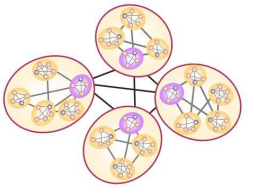
- **Key idea:**
  - Greedy layer-wise training
  - Pre-training + fine tuning
  - Restricted Boltzmann Machine
  - Contrastive divergence

단계별 발달 과정 도입



The importance of topology



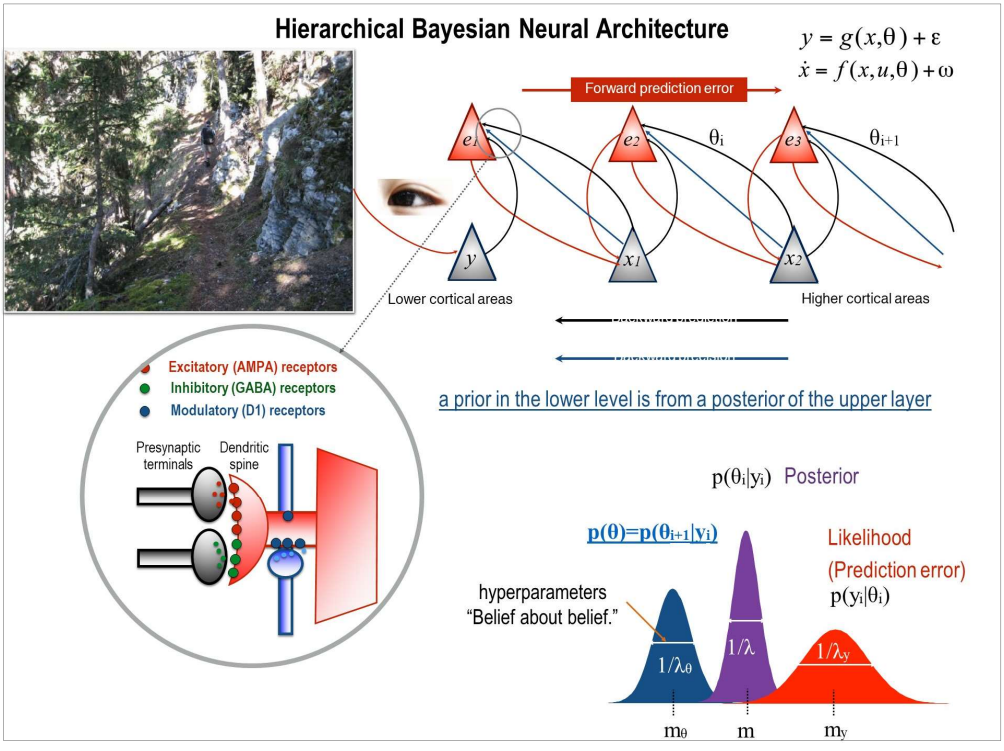


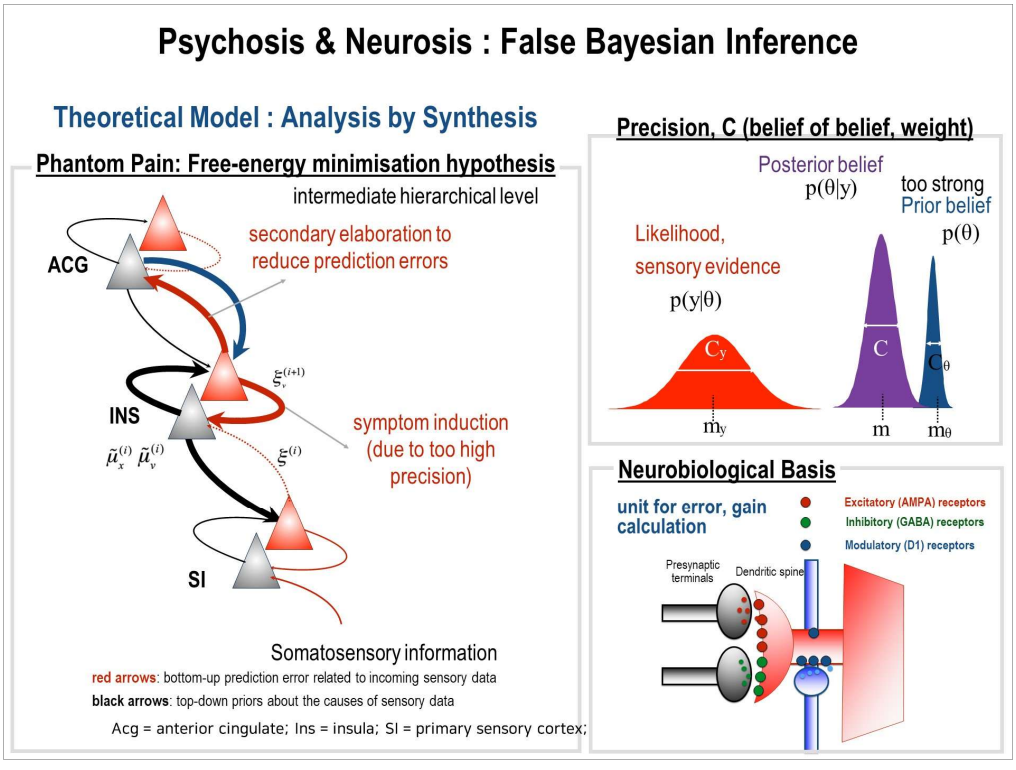
Park and Friston, SCIENCE, 2013

Generative Pre-trained Transformer

ATTENTION

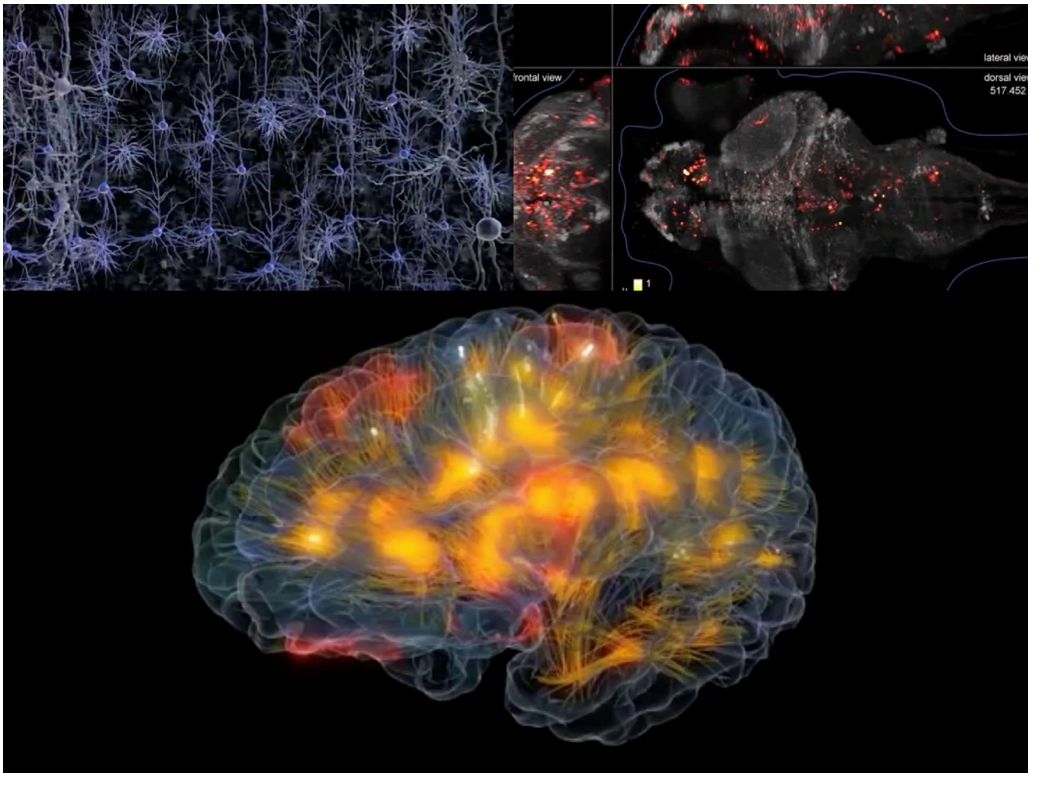
PRECISON

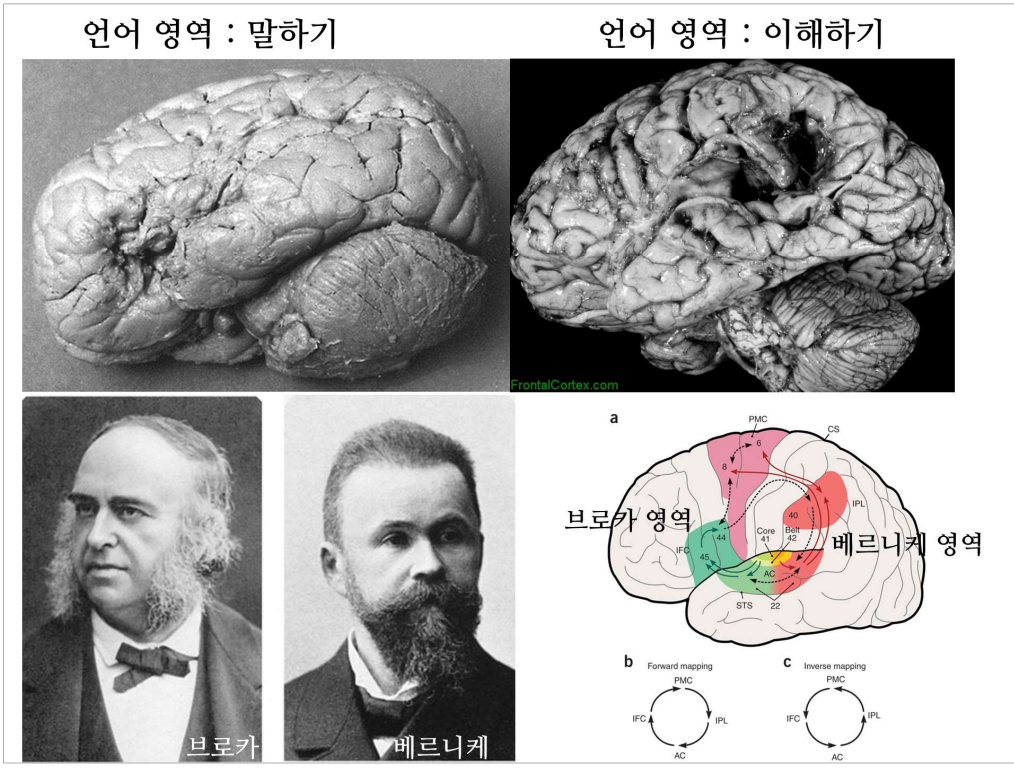

Hierarchical Bayesian Neural Architecture
$y = g(x,\theta) + \varepsilon$
$\dot{x} = f(x,u,\theta) + \omega$
Forward prediction error
e_1
e_2
θ_i
e_3
θ_{i+1}
y
x_1
x_2
Lower cortical areas
Higher cortical areas
Excitatory (AMPA) receptors
Inhibitory (GABA) receptors
Modulatory (D1) receptors
Presynaptic terminals
Dendritic spine
a prior in the lower level is from a posterior of the upper layer
$p(\theta_i|y_i)$ Posterior
$p(\theta)=p(\theta_{i+1}|y_i)$
Likelihood
(Prediction error)
$p(y_i|\theta_i)$
hyperparameters
"Belief about belief."
$1/\lambda_\theta$
$1/\lambda$
$1/\lambda_y$
m_θ
m
m_y


Psychosis & Neurosis : False Bayesian Inference
Theoretical Model : Analysis by Synthesis
Phantom Pain: Free-energy minimisation hypothesis
intermediate hierarchical level
ACG
secondary elaboration to reduce prediction errors
INS
SI
symptom induction (due to too high precision)
Somatosensory information
red arrows: bottom-up prediction error related to incoming sensory data
black arrows: top-down priors about the causes of sensory data
Acg = anterior cingulate; Ins = insula; SI = primary sensory cortex;
Precision, C (belief of belief, weight)
Posterior belief
$p(\theta|y)$
too strong
Prior belief
$p(\theta)$
Likelihood,
sensory evidence
$p(y|\theta)$
C_y
C
C_θ
m_y
m
m_θ
Neurobiological Basis
unit for error, gain calculation
Excitatory (AMPA) receptors
Inhibitory (GABA) receptors
Modulatory (D1) receptors
Presynaptic terminals
Dendritic spine

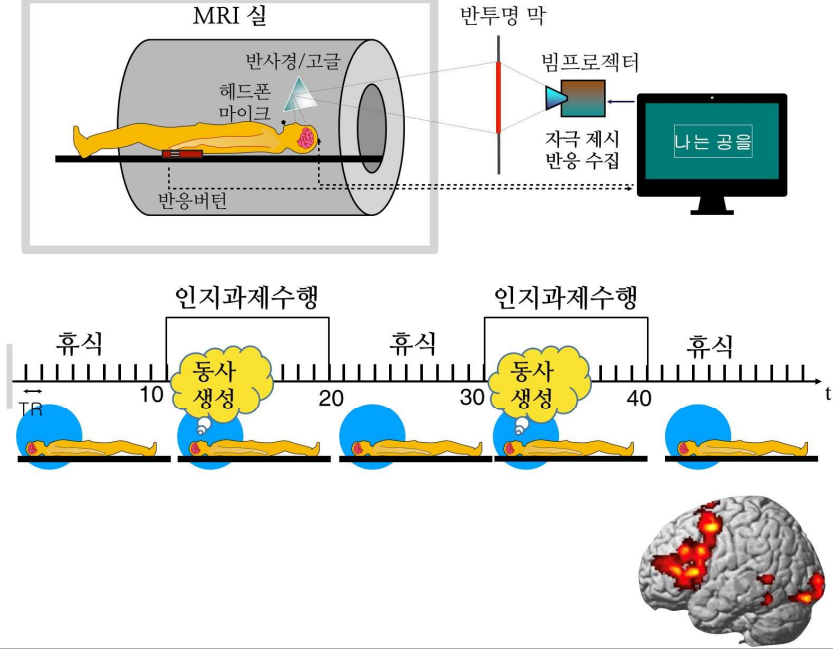
ISSUES

ChatGPT proves the brain principle is relatively simple.
Other Cognitive Faculty is not impossible…



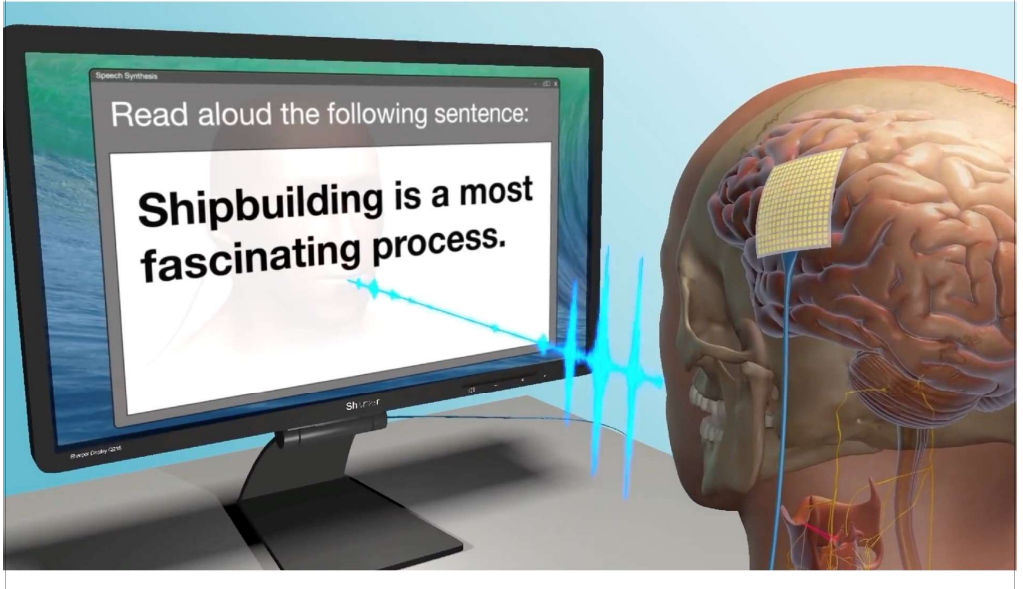

언어 영역 : 말하기
언어 영역 : 이해하기
FrontalCortex.com
브로카
베르니케
a
PMC
CS
6
8
IPL
40
Core 41
Belt 42
44
IFG
45
AC
STS
22
브로카 영역
베르니케 영역
b
Forward mapping
PMC
IFG
IPL
AC
c
Inverse mapping
PMC
IFG
IPL
AC

동사 생성 뇌 활성

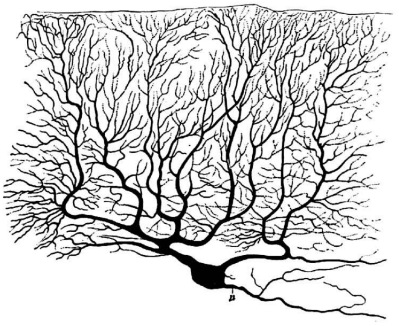

MRI 실
반사경/고글
헤드폰
마이크
반응버튼
반투명 막
빔프로젝터
자극 제시
반응 수집
나는 공을
인지과제수행
인지과제수행
휴식
휴식
휴식
동사
생성
동사
생성
TR
10
20
30
40
t

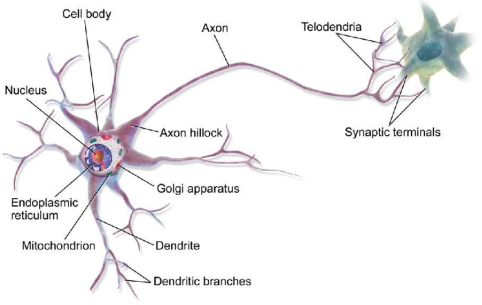
뇌과학: 정신 작용은 물리적 측정 가능한 뇌 활동으로 환원

EBS 다큐멘타리

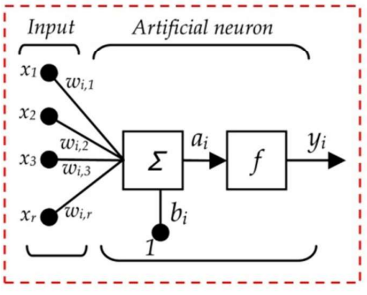


Neurons and Neural Network





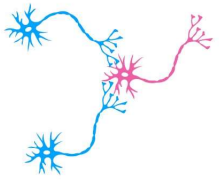
Neural Network



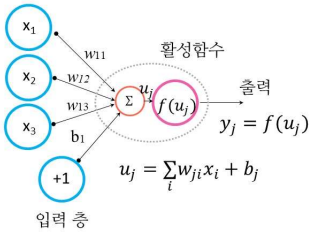
$$y_i = f\left(\sum_j w_{ij}x_j + b_i\right)$$

f: activation function

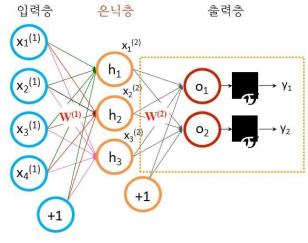
뉴론과 시냅스



퍼셉트론 (~1957)



다층퍼셉트론 (~1969)



대표적 활성함수 f

Logistic (sigmoid)

$$f(x) = \frac{1}{1+\exp(-x)}$$

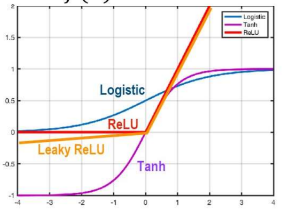
$$f'(x) = f(x)(1-f(x))$$

Rectified Linear Unit (ReLU)

$$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$$

$$f'(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$$

$f(u)$: 활성함수



출력층 분류: softmax 함수 σ

K 개의 출력층 노드 수: 클래스 수

$$o_k = \sum_j w_{kj}^{(2)} x_j^{(2)} + b_k^{(2)}$$

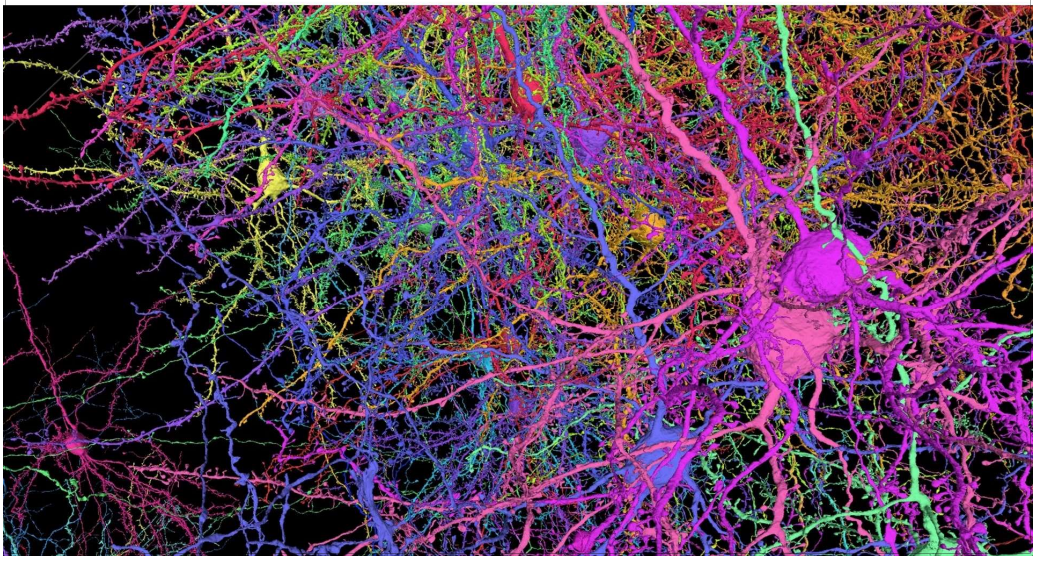
$$y_k = \sigma(o_k) = \frac{e^{o_k}}{\sum_{k'=1}^{K} e^{o_{k'}}}$$

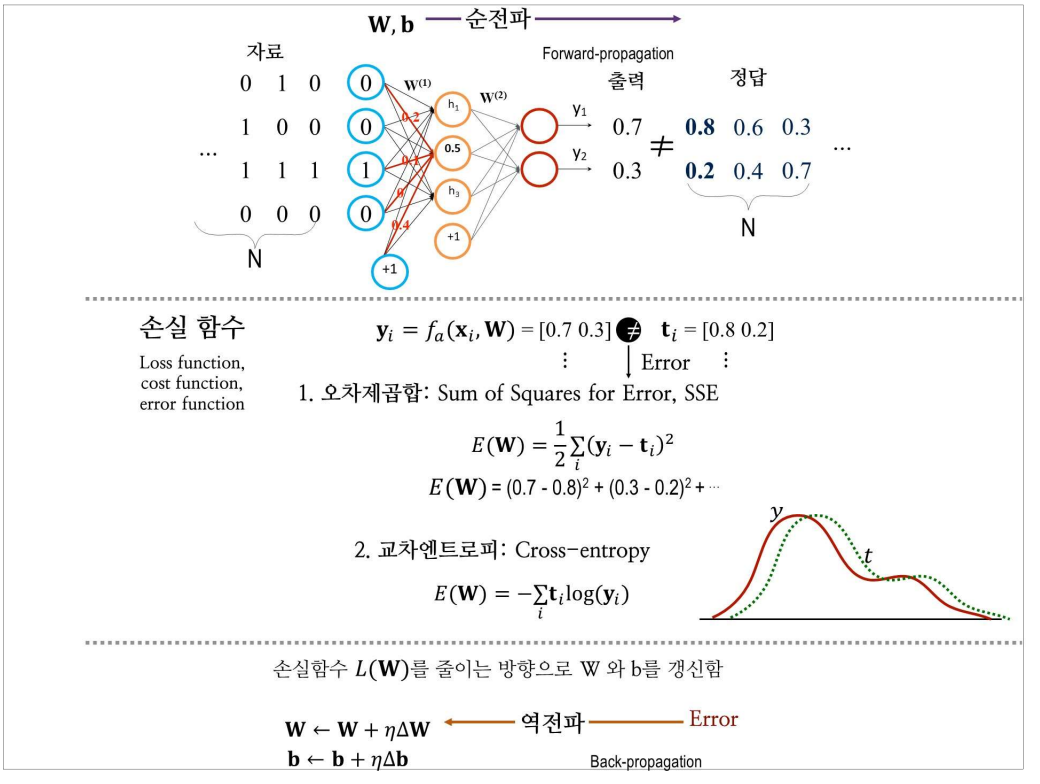
입력 x 가 주어졌을때
그 클래스가 k 일 확률

$$P(y = k|x) = y_k$$

Optimism. Everything is from connection. Human mind is impossible in some day…

https://www.technologyreview.com/2021/08/02/1030453/microns-connections-in-a-mouse-brain/





MIT Technology Review

Featured Topics Newsletters Events Podcasts SIGN IN SUBSCRIBE

ARTIFICIAL INTELLIGENCE

How to solve AI's inequality problem

New digital technologies are exacerbating inequality. Here's how scientists creating AI can make better choices.

By David Rotman April 19, 2022

IAN GRANDJEAN

The economy is being transformed by digital technologies, especially in artificial intelligence, that are rapidly changing how we live and work. But this transformation poses a troubling puzzle: these technologies haven't done much to grow the economy, even as income inequality worsens. Productivity growth, which economists consider essential to improving living standards, has largely been sluggish since at least the mid-2000s in many countries.

Why are these technologies failing to produce more economic growth? Why aren't they fueling more widespread prosperity? To get at an answer, some leading economists and policy experts are looking more closely at how we invent and deploy AI and automation—and identifying ways we can make better choices.

POPULAR

Geoffrey Hinton tells us why he's now scared of the tech he helped build

Will Douglas Heaven

ChatGPT is going to change education, not destroy it

A conversation with Eliza

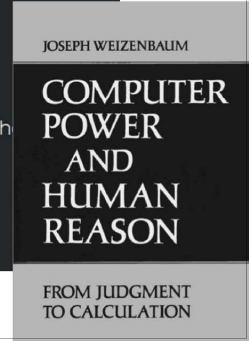
```
Welcome to
                EEEEEE  LL      IIII   ZZZZZZ   AAAAA
                EE      LL       II       ZZ   AA   AA
                EEEEE   LL       II     ZZZ    AAAAAAA
                EE      LL       II    ZZ      AA   AA
                EEEEEE  LLLLLL  IIII ZZZZZZ    AA   AA

  Eliza is a mock Rogerian psychotherapist.
  The original program was described by Joseph Weizenbaum in 1966.
  This implementation by Norbert Landsteiner 2005.


ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU:   Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come h
YOU:   He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU:   It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:
```

Joseph Weizenbaum

1976년



'Godfather of AI' discusses dangers the developing technologies pose to society

PBS NEWS HOUR

SEJ · SEO

Top 5 Ethical Concerns Raised By AI Pioneer Geoffrey Hinton

AI pioneer Geoffrey Hinton warns of machines surpassing human intelligence, raises ethical concerns, and departs Google to promote responsible AI development.

In light of his observations of new large language models like GPT-4, Hinton cautions about several key issues:

1. **Machines surpassing human intelligence**: Hinton believes AI systems like GPT-4 are on track to be much smarter than initially anticipated, potentially possessing better learning algorithms than humans.
2. **Risks of AI chatbots being exploited by "bad actors"**: Hinton highlights the dangers of using intelligent chatbots to spread misinformation, manipulate electorates, and create powerful spambots.
3. **Few-shot learning capabilities**: AI models can learn new tasks with just a few examples, enabling machines to acquire new skills at a rate comparable to, or even surpass, that of humans.
4. **Existential risk posed by AI systems**: Hinton warns about scenarios in which AI systems create their own subgoals and strive for more power, surpassing human knowledge accumulation and sharing capabilities.
5. **Impact on job markets**: AI and automation can displace jobs in certain industries, with manufacturing, agriculture, and healthcare being particularly affected.

CHRISTIAN PERSPECTIVE

Generative Pre-trained Transformer

THANK YOU