Interpretability, Responsibility and Controllability of Human Behaviors

Xiaohong Wan

Beijing Normal University

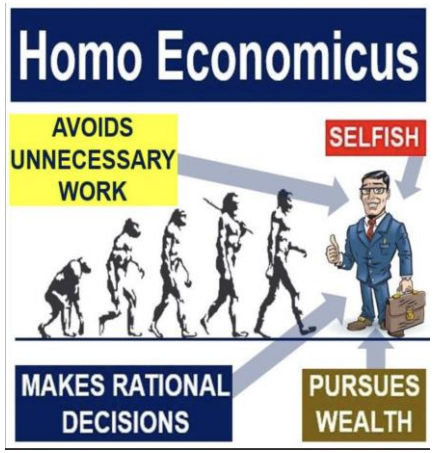
Email: xhwan@bnu.edu.cn

2022.10.24

“The Rationality of man” is the key basis of Building Human Societies

“The difference between the reason of man and the instict of the beast is this, that the beast does but know, but the man knows that he knows.”
--John Donne

Economy



Law

Politics

Insanity of Man: automatism out of control

Man who killed wife while dreaming is freed

November 20, 2009 -- Updated 1822 GMT (0222 HKT)

STORY HIGHLIGHTS

- Prosecutors described trial involving Brian Thomas as a "unique case"
- Thomas killed wife during a bad dream while pair was vacationing in 2008
- Judge described Thomas as a "decent man and devoted husband"

London, England (CNN) -- A British man who strangled his wife in his sleep while dreaming that she was an intruder walked free from court Friday after the case against him was withdrawn, prosecutors said.

The UK's Crown Prosecution Service requested that the case against Brian Thomas, who killed wife Christine while they were on vacation in 2008, be dropped due to a "unique set of circumstances."

Thomas, 59, of Neath, South Wales, had been on trial for murder at Swansea Crown Court, with prosecutors seeking a verdict of not guilty by reason of insanity that would have resulted in psychiatric custody.

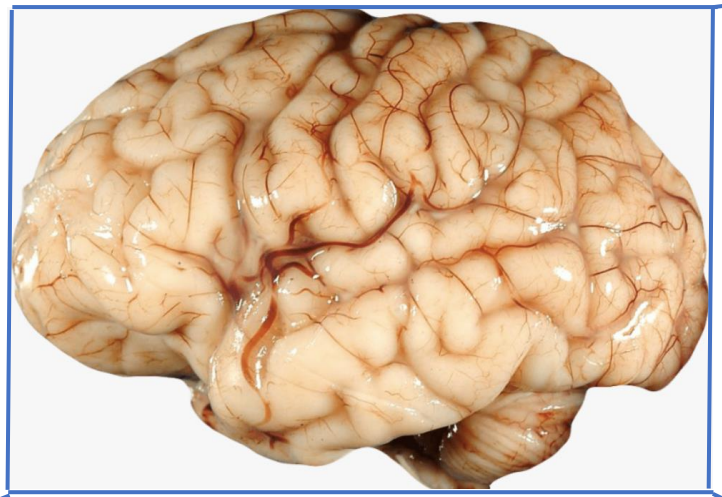
But they said a closer study of evidence led them to believe the defendant should be released without further action.

"We have duty to keep cases under continuous review, and following expert evidence from a psychiatrist it was suggested no useful purpose would be served by Mr Thomas being detained and treated in a psychiatric hospital," prosecutor Iwan Jenkins said in a statement.

Automatism implies that the accused person had no control over his actions, that he acted like a runaway machine

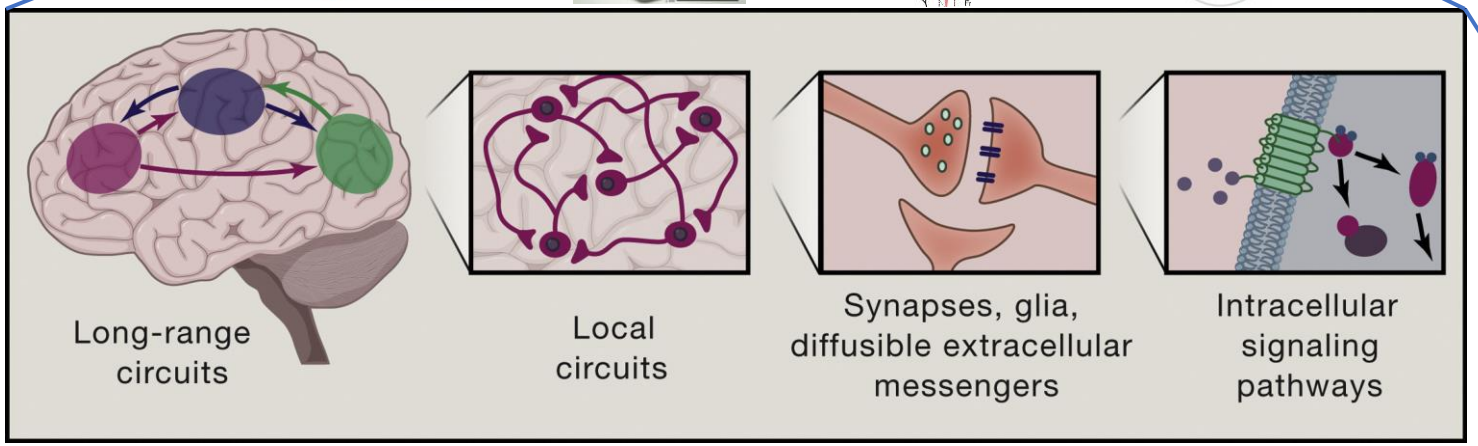
Responsibility entails controllability, but not interpretability

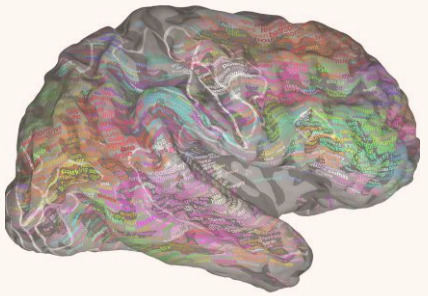
Neuroscience: Interpreting the relatonships between brain activities and behaviors



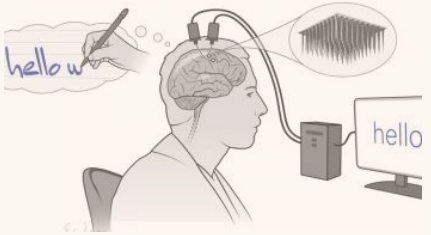
Decoding & Intervening

Observation

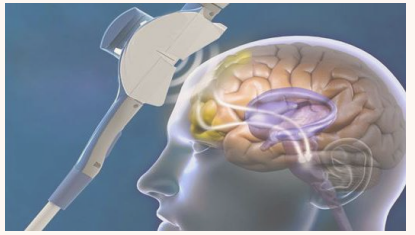




brain dictonary



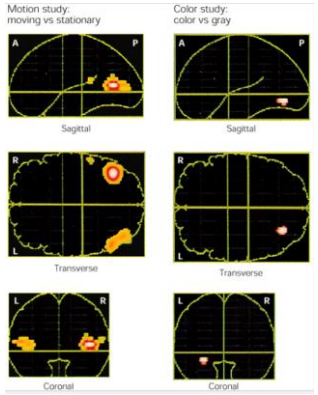
brain speech

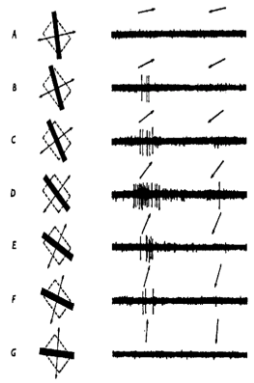


causal behaviors

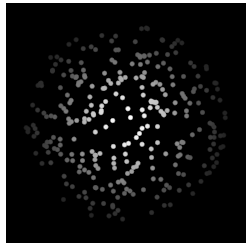
An example of relationship between neural activities and perception

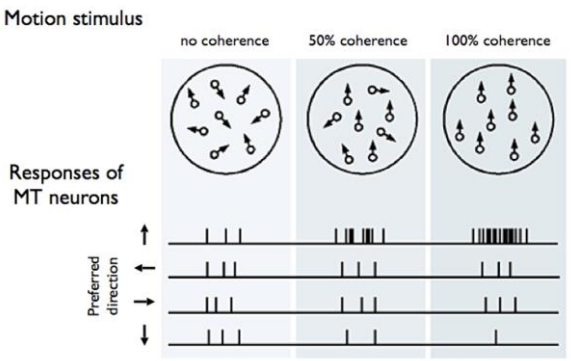
Motion-selective area (MT)





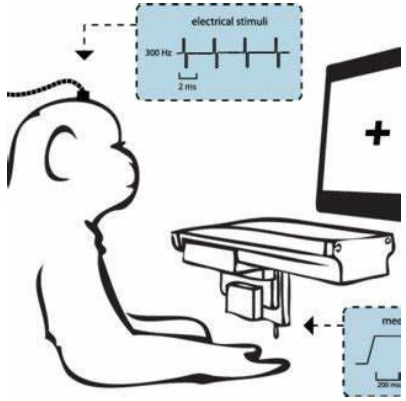
Zeki et al., 1987

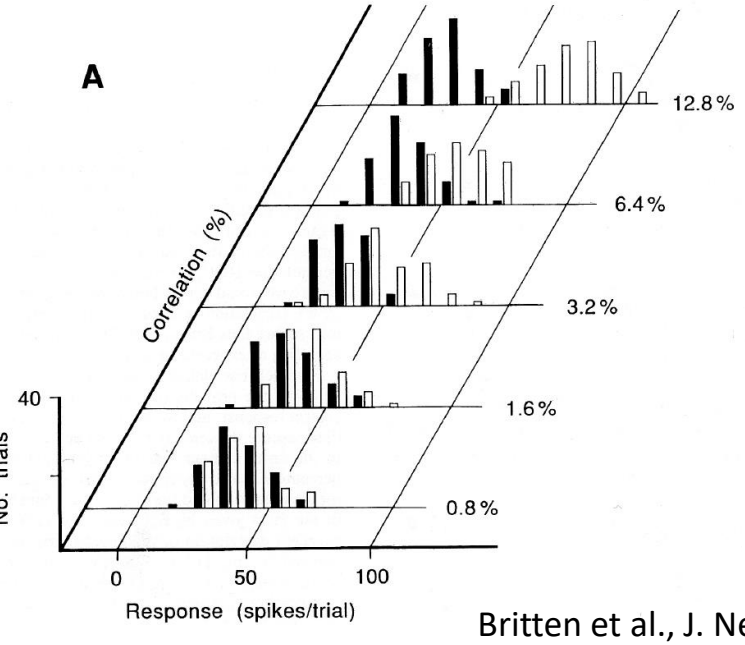


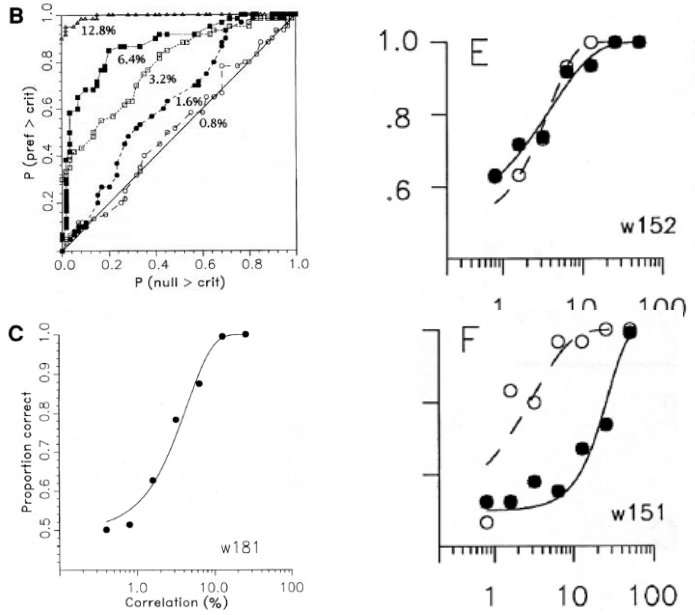


Albright et al., J. Neurophysiol. (1984)

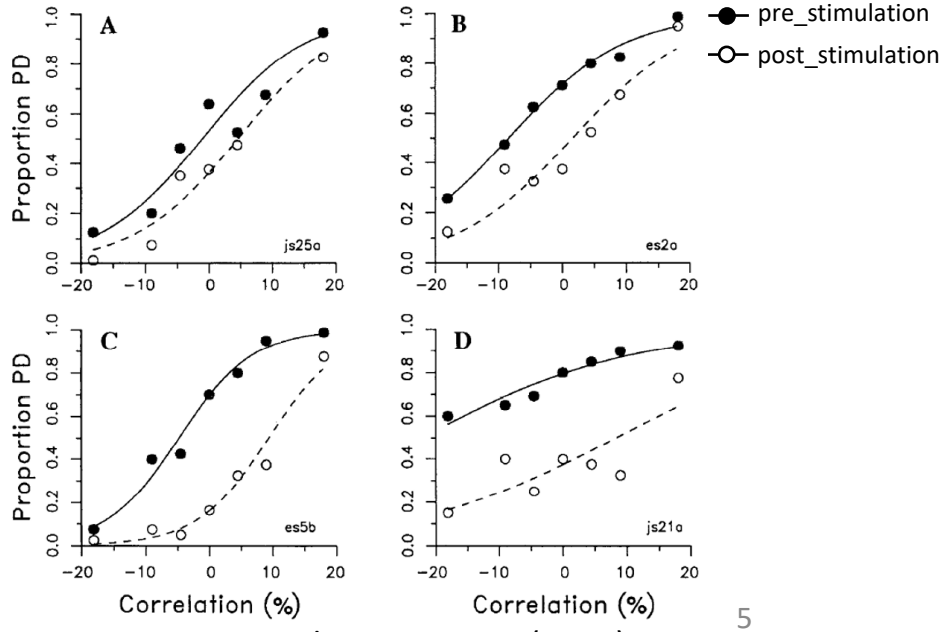
microstimulation





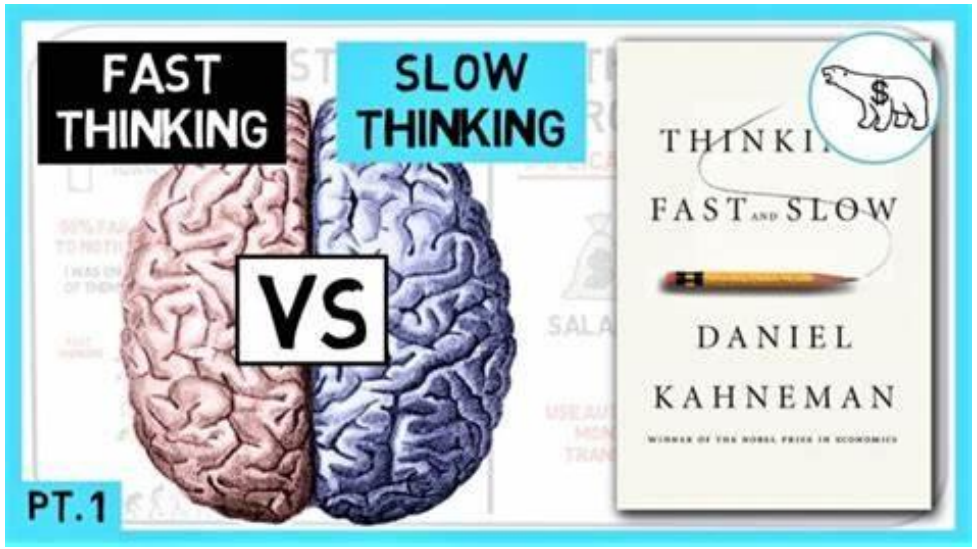


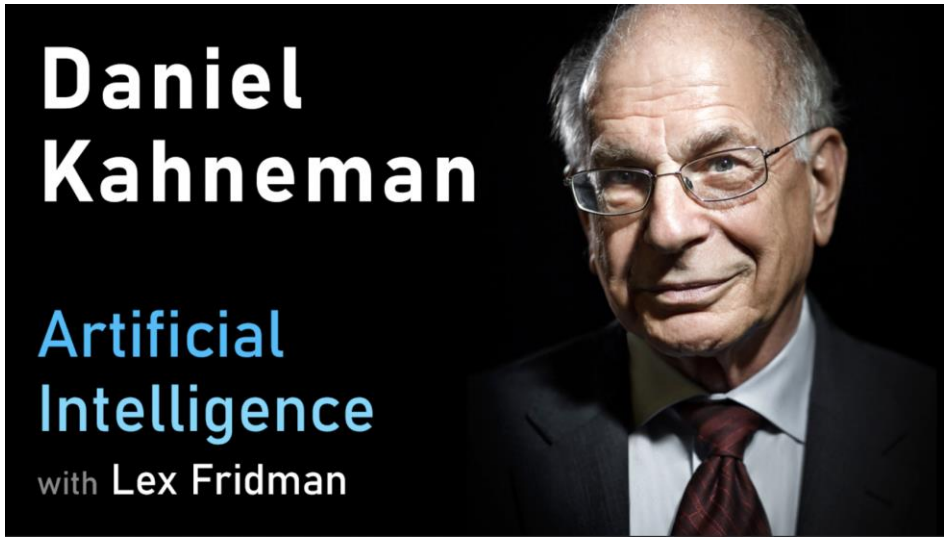
Britten et al., J. Neurosci. (1992)

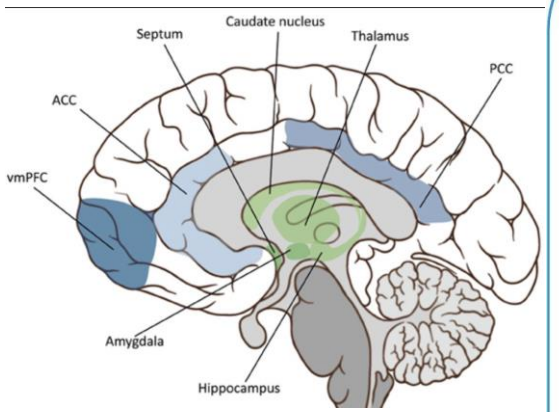


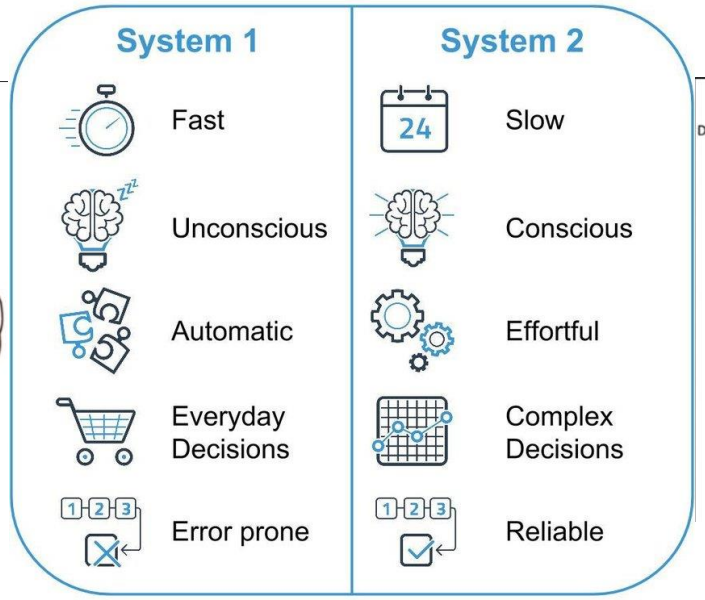
Britten et al., J. Neurosci. (1992)

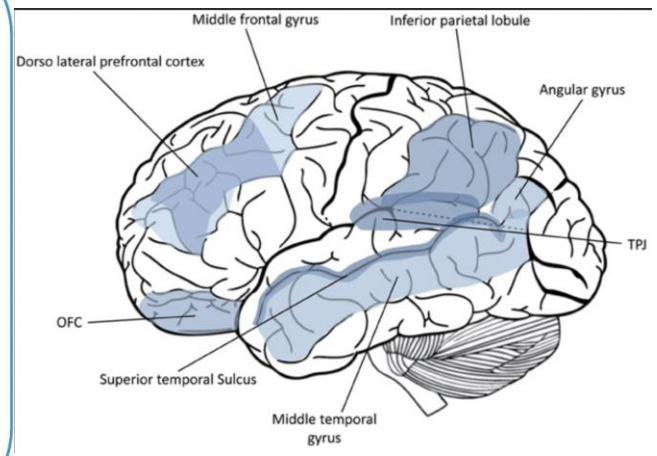
1. “The Rationality of mans” have two parallel systems





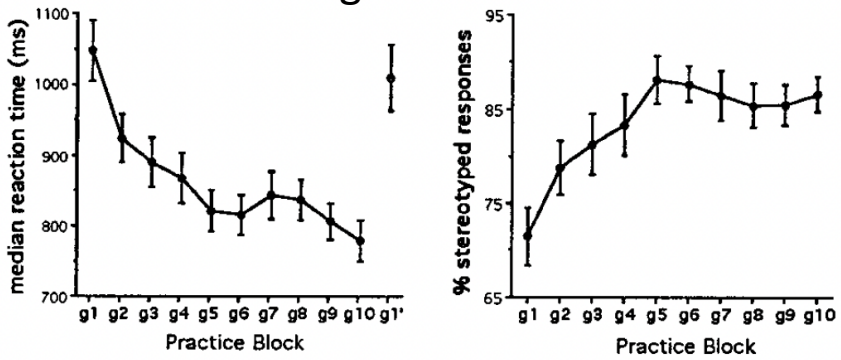


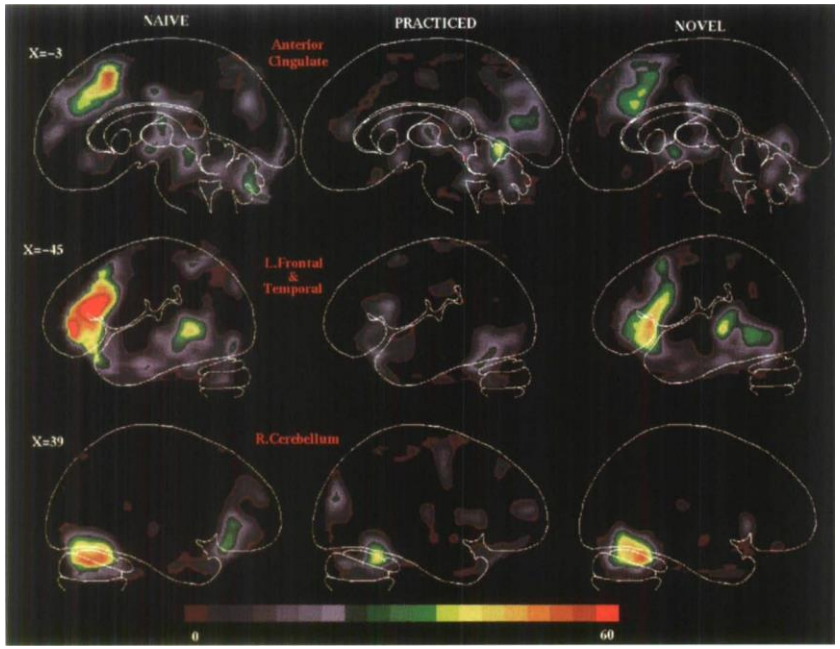




Practices shift human brain activations from system 2 to system 1

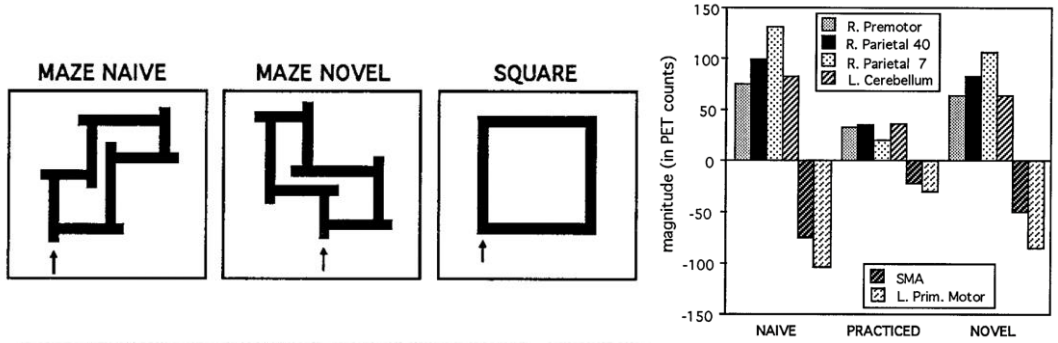
Verbal generation task

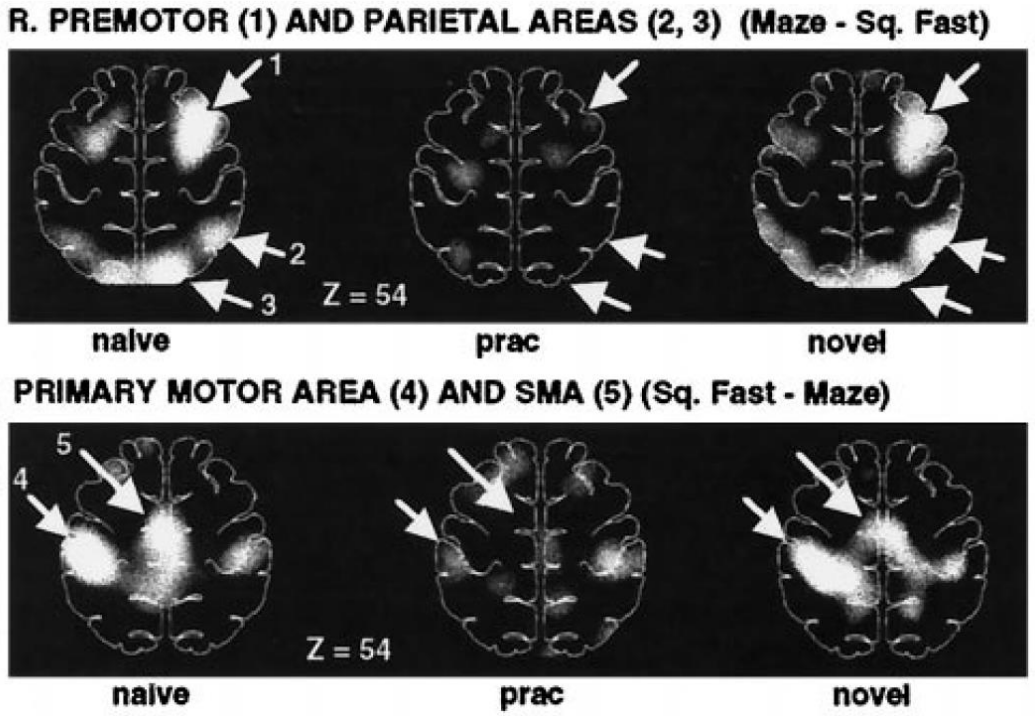




PET

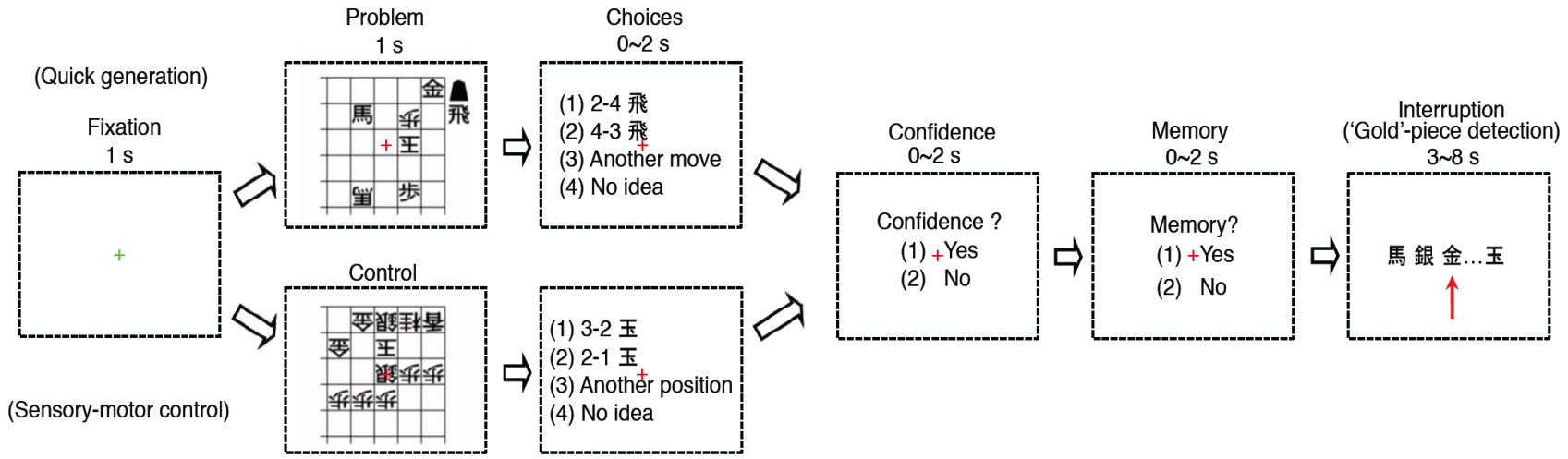
Raichle et al., Cereb. Cortex 1994

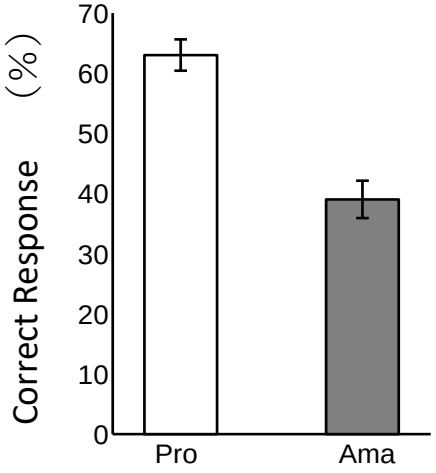


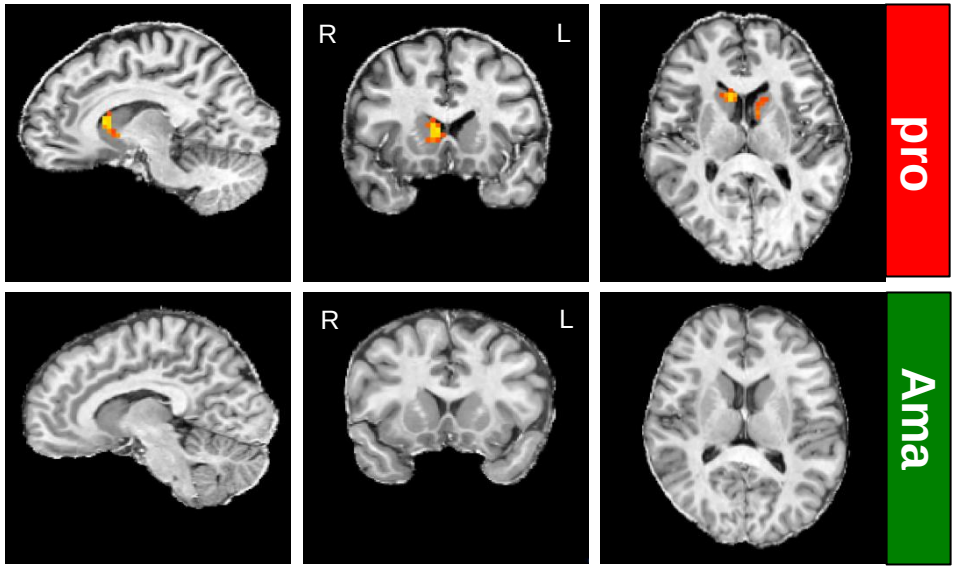


Petersen et al., PNAS 1998

Experts' intuitive decisions rely on caudate nucleus in straitum



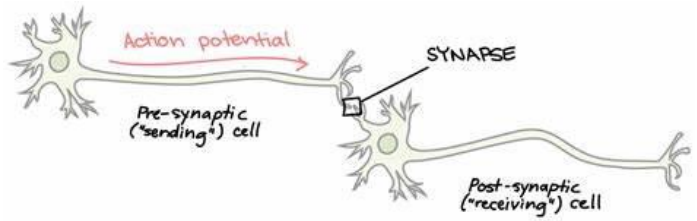




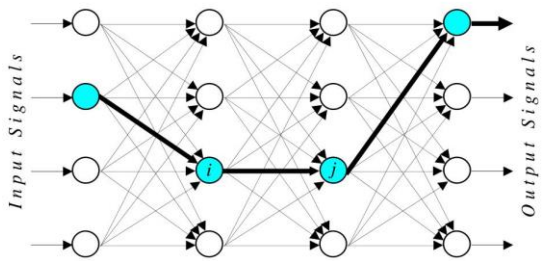
Wan et al. *Science* (2011); Wan et al. *J. Neurosci.* (2012)

Associative learning: formations of habits and intuitions

1. Hebbian learning: fire together, wire together

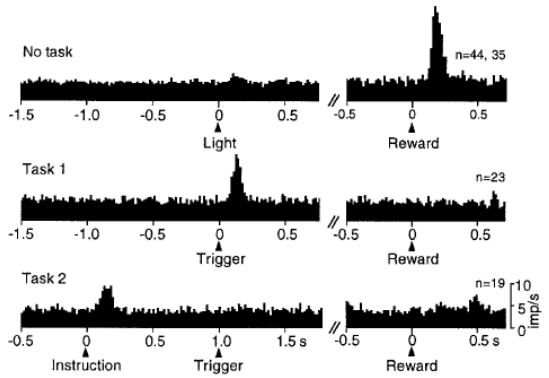


Hebbian learning in a neural network

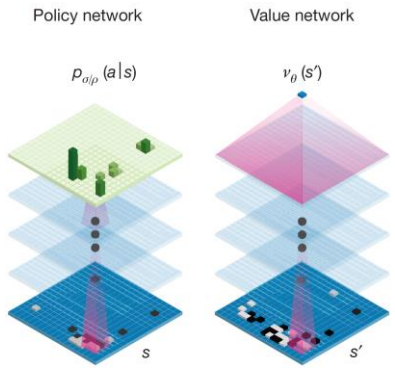


$$\Delta w_{ij}(p) = \alpha \; y_j(p) \; x_i(p)$$

2. Reinforcement learning



$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t) \right]$$



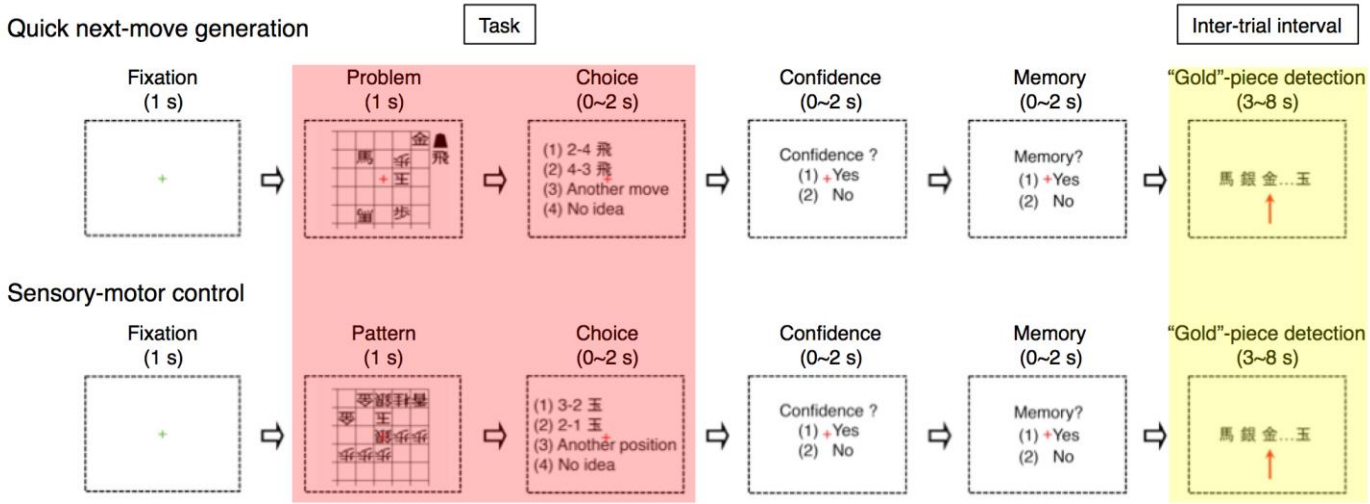
Heuristics

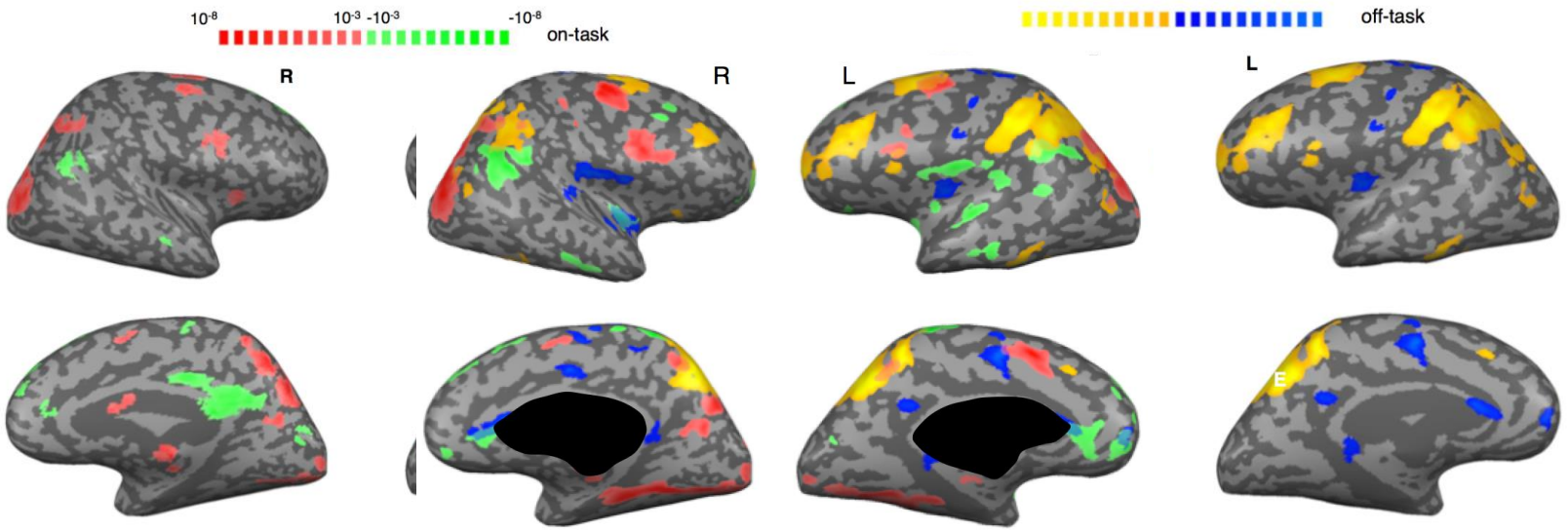
Habits

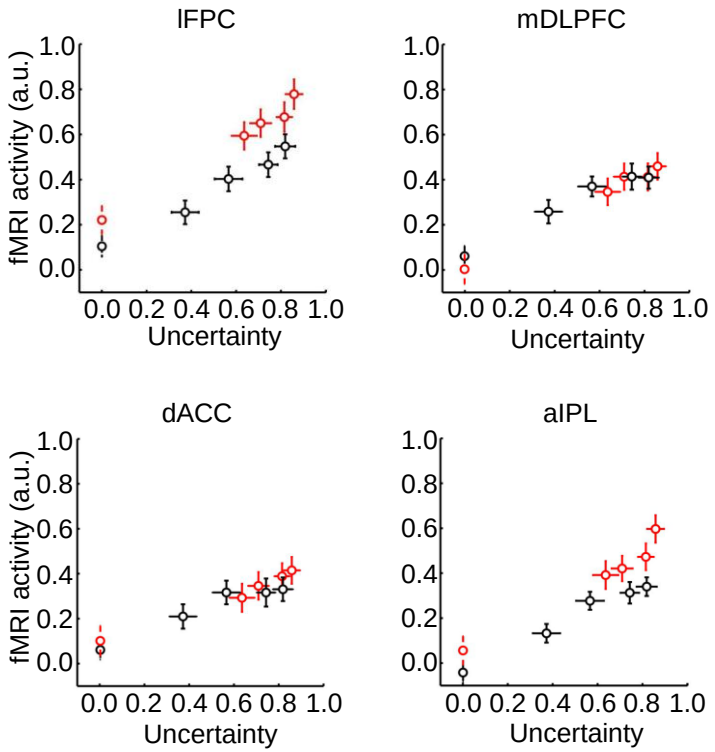
Intuitions

automatic
spurious corrlations
not necessarily causality

A meta-level neural system monitors the outcomes of system 1

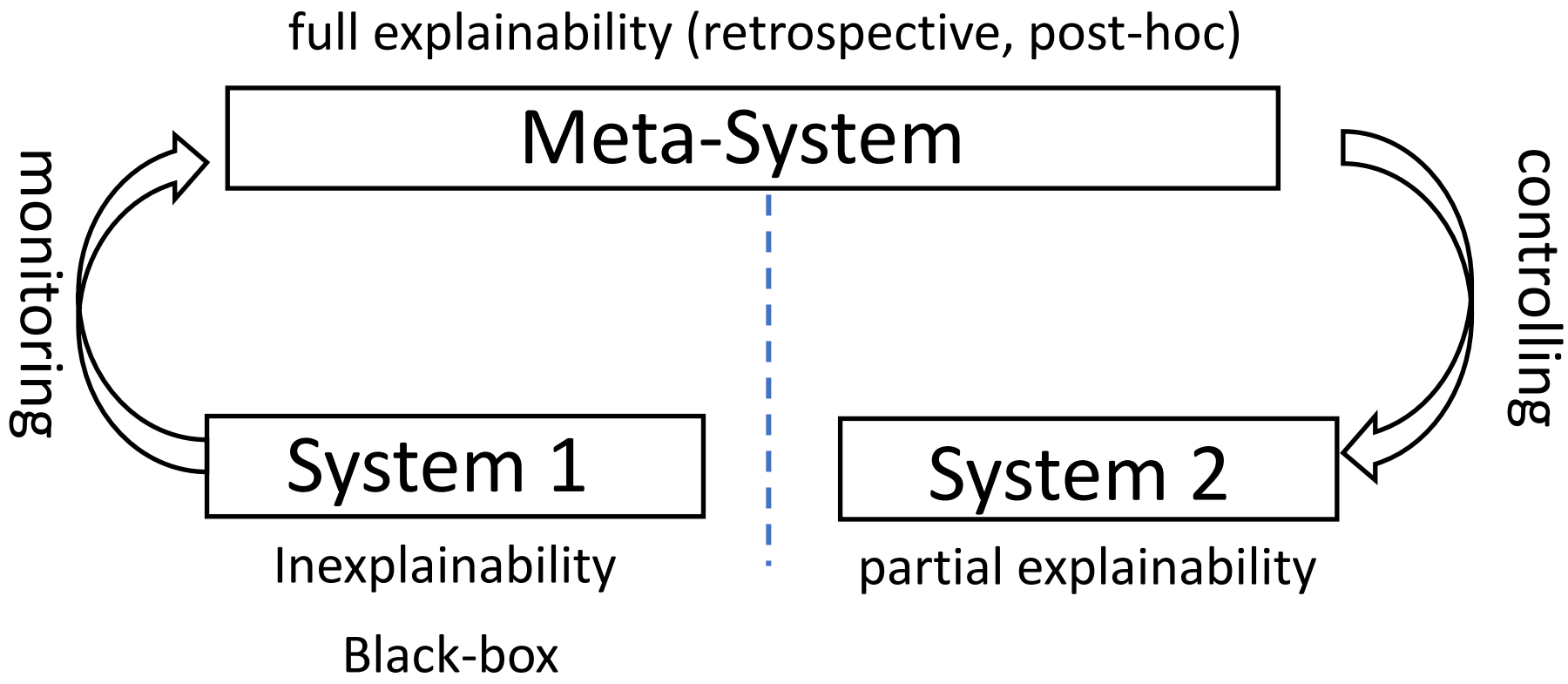






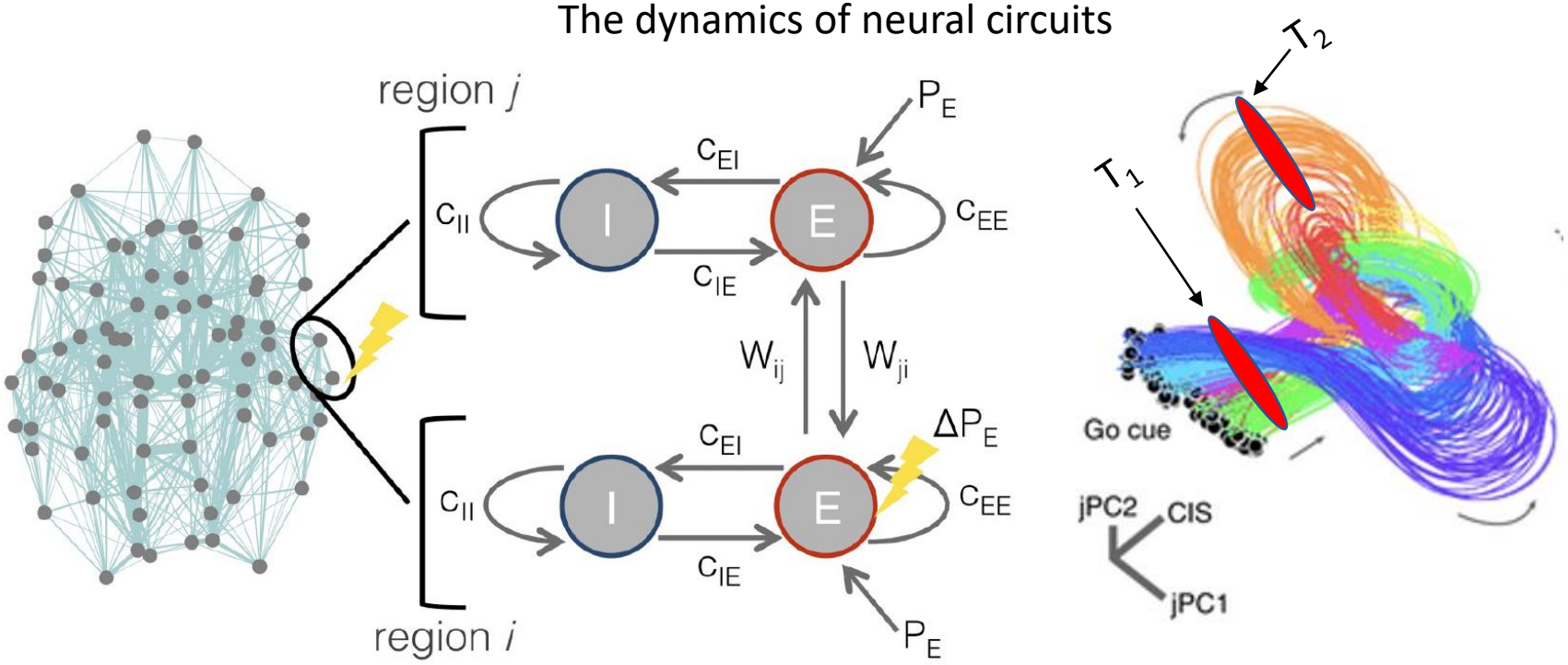
Wan et al. eNeuro (2016)

A meta-system for monitoring, controlling and explainations



2. The neural dynamics shape post-hoc explainations

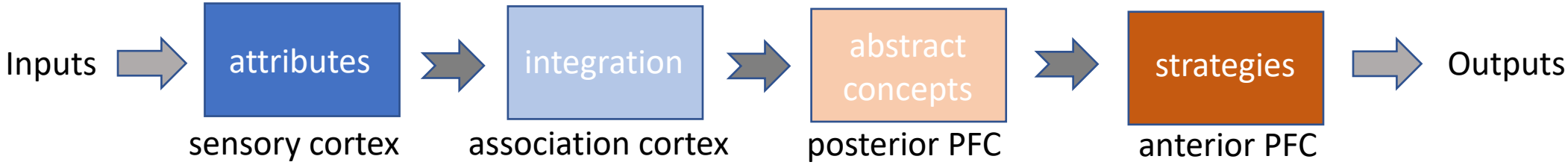
The dynamics of neural circuits



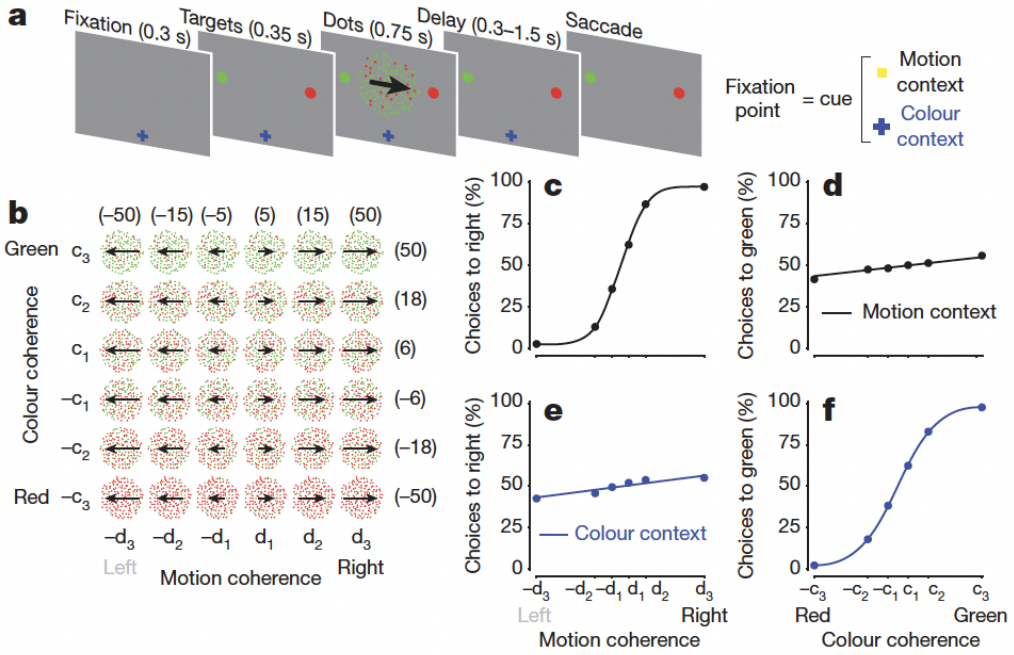
Confirmation bias, Primacy effects, Recency effects, and so on

3. The entanglemts of high-dimentional Neural manifolds in prefrontal cortex

Hierachy of neural representations in human brain

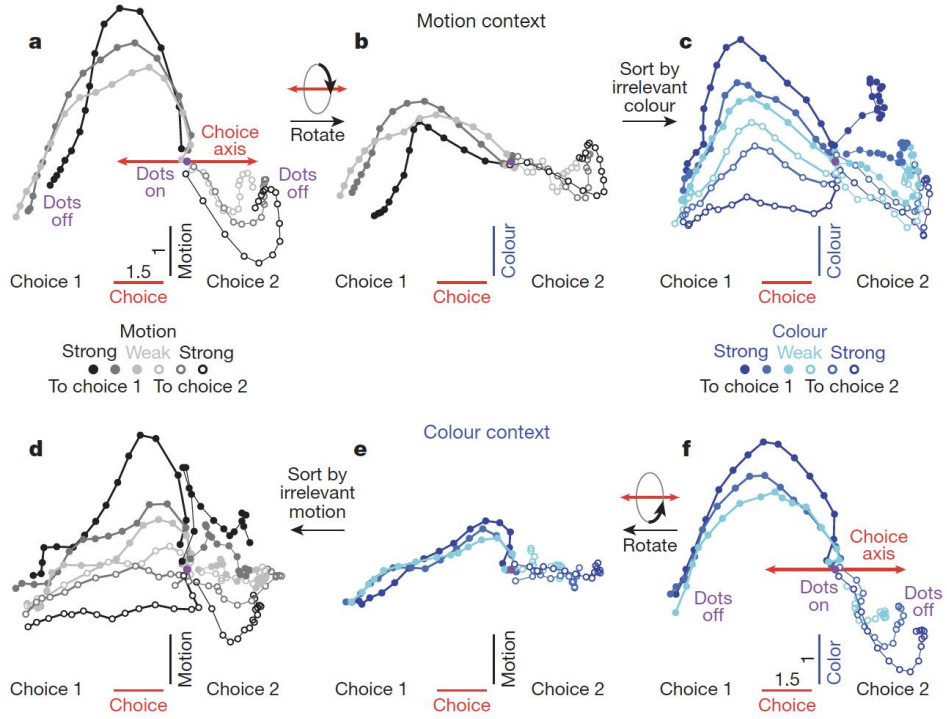


Context-dependent tasks



Mante et al., Nature (2013)

Context-dependent disentanglement in prefrontal cortex



Insights from Neuroscience: explanations of neural processes

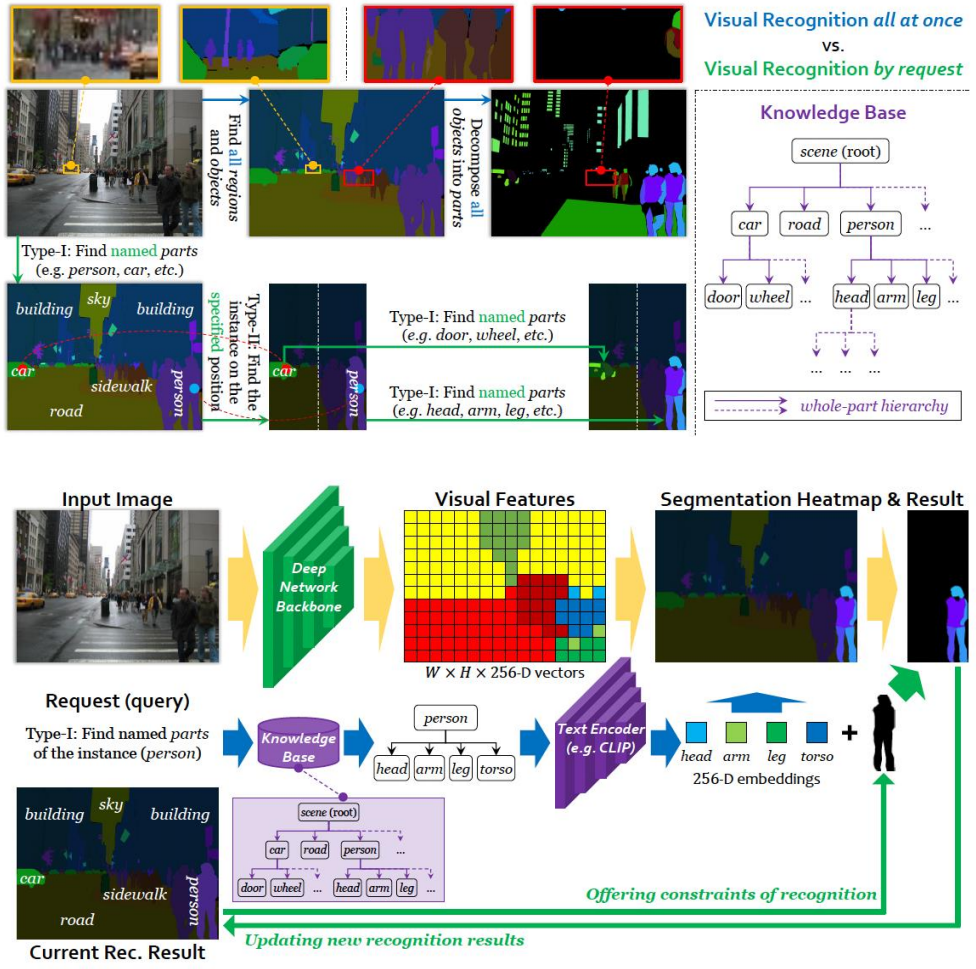
1. Many processes are automatic and black-box in humans.
   most of these automatic processes are monitored by the higher-order neural system.
   (meta-cognition)

2. The post-hoc explanations of neural processes are deviated from the originals, due to the dynamics of neural circuits.
   False memory, Illusions, cognitive biases

3. The high-dimensional neural representations are disentangled when the goal is driven, but not separately and hierachically.
   (like transformer needs prompting)

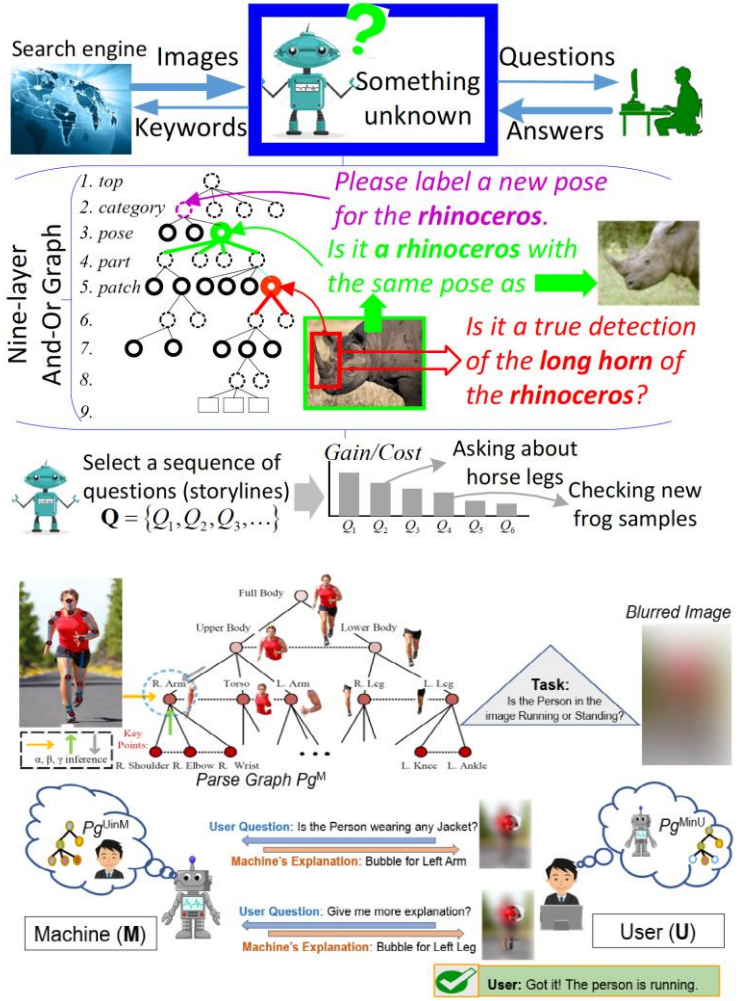
1. Query-based Recognition and Reasoning

Pixel-level semantical recognitions



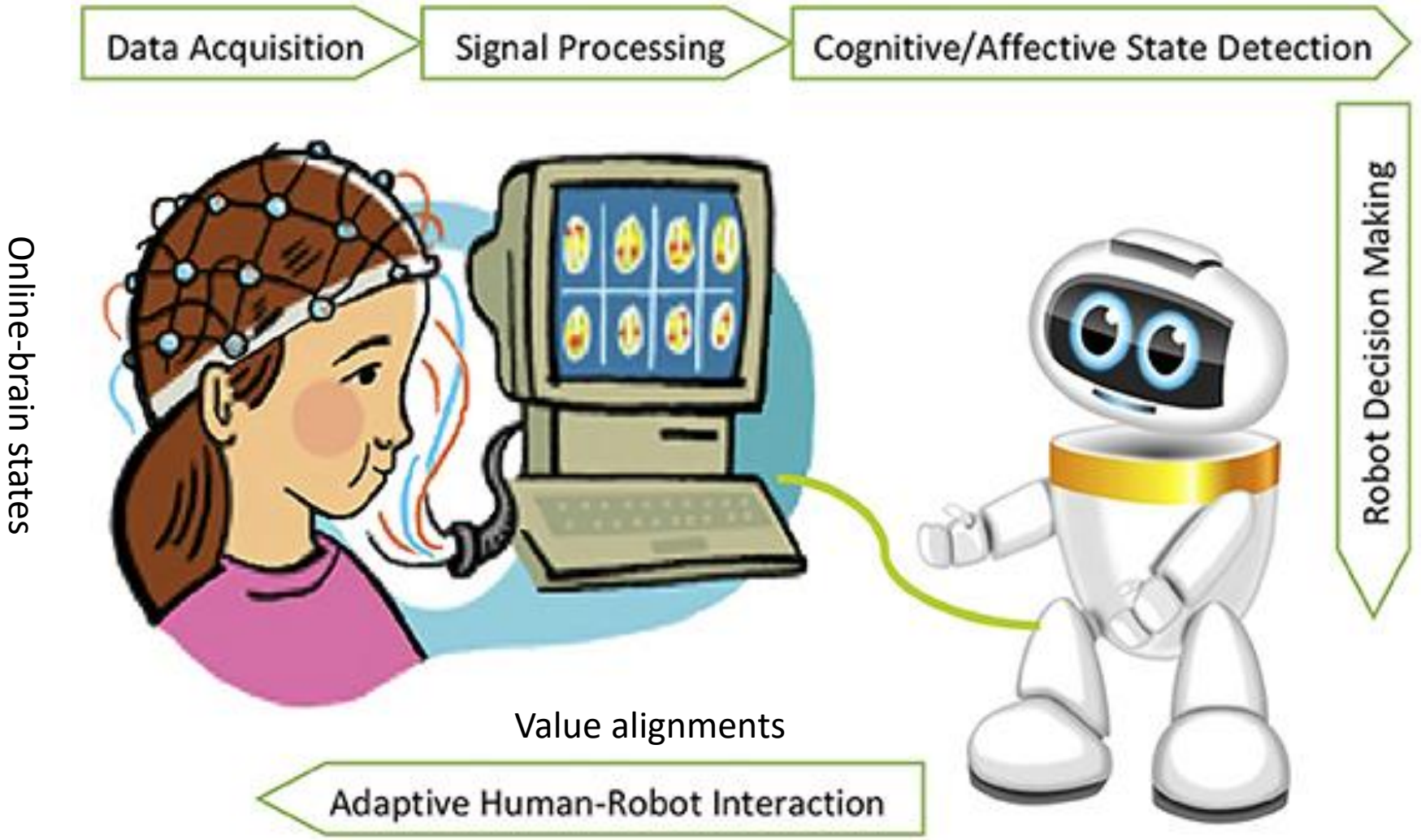
Tang et al., Visual recogniton by request. ArXiv (2022)

X-ToM: Explaining with theory of mind



Akula et al., . ArXiv (2019)

2. BCI-supported Human-AI hybrid Incorporation



Thanks for your attentions