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 instinct of the beast is this, that the beast does but know, but the man knows that he knows.”
--John Donne
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 relationships between brain activities and behaviors

Observation

Decoding & Intervening

Long-range circuits

Local circuits

Synapses, glia, diffusible extracellular messengers

Intracellular signaling pathways

brain dictionary

brain speech

causal behaviors
An example of relationship between neural activities and perception

Motion-selective area (MT)

Zeki et al., 1987


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

System 1
- Fast
- Unconscious
- Automatic
- Everyday Decisions
- Error prone

System 2
- Slow
- Conscious
- Effortful
- Complex Decisions
- Reliable
Practices shift human brain activations from system 2 to system 1

**Verbal generation task**

Raichle et al., Cereb. Cortex 1994

Petersen et al., PNAS 1998
Experts’ intuitive decisions rely on caudate nucleus in straitum

Associative learning: formations of habits and intuitions

1. Hebbian learning: fire together, wire together

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 correlations
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 explanations

Inexplainability
Black-box

full explainability (retrospective, post-hoc)

Meta-System

System 1

System 2

partial explainability
2. The neural dynamics shape post-hoc explanations

The dynamics of neural circuits

Confirmation bias, Primacy effects, Recency effects, and so on
3. The entanglements of high-dimensional Neural manifolds in prefrontal cortex

Hierachy of neural representations in human brain

- Inputs: attributes -> integration -> abstract concepts -> strategies -> Outputs
  - sensory cortex -> association cortex -> posterior PFC -> anterior PFC

Context-dependent tasks

Context-dependent disentanglement in prefrontal cortex

Mante et al., Nature (2013)
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 hierarchically. (like transformer needs prompting)
1. Query-based Recognition and Reasoning

Pixel-level semantical recognitions

Input Image

Visual Features

Segmentation Heatmap & Result

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

X-ToM: Explaining with theory of mind

Search engine

Images

Questions

Keywords

Answers

Please label a new pose for the rhinoceros.
Is it a rhinoceros with the same pose as

Is it a true detection of the long horn of the rhinoceros?

Select a sequence of questions (storylines)

Gain/Cost

Asking about

horse legs

Checking new frog samples

Akula et al., ArXiv (2019)
2. BCI-supported Human-AI hybrid Incorporation

- Data Acquisition
- Signal Processing
- Cognitive/Affective State Detection

Online-brain states
Value alignments
Adaptive Human-Robot Interaction
Robot Decision Making
Thanks for your attentions