Centre for Cognitive Science

Computational cognitive neuroscience:
11. Agency, theory of mind, and consciousness

Lubica Beňušková
Centre for Cognitive Science, FMFI
Comenius University in Bratislava

"Cogito ergo sum"

- When I think, there must be something doing the thinking. René Descartes (1596–1650) calls this the mind or soul. The origin of the modern concept of consciousness is attributed to John Locke and his Essay Concerning Human Understanding, published in 1690.

- The English word "conscious" derives from the Latin "conscius" (con- 'together' and scio 'to know') —it meant "knowing with", in other words, "having joint or common knowledge with another".

- There were, however, many occurrences in Latin writings of the phrase conscio sibi, which translates literally as "knowing with oneself", or "sharing knowledge with oneself about something".

- 75% of the Latin literature 90–40 BC was written by a single man: Marcus Tullius Cicero who created a Latin philosophical vocabulary with neologisms such as evidentia, humanitas, qualitas, quantitas, essentia, and others.

The sense of agency

- The "sense of agency" (SA) (or sense of control) refers to the subjective awareness that one is initiating, executing, and controlling one's own volitional actions in the world.
  - v Slovenčine žiaľ neexistuje uspokojivý preklad: možno čosi ako 'ja som ten agent, čo riadi túto akciu'.

- It is the pre-reflective awareness that it is **I who is controlling** bodily movement(s) or thinking thoughts.

- In normal non-pathological experience, the SA is tightly integrated with one's "sense of ownership" (SO).

The sense of agency versus sense of ownership

- The sense of ownership (SO) is a pre-reflective awareness that one is the **owner** of an action, movement or thought, it is s/he who executes them.

- The sense of agency (SA), or sense of **control**, is the subjective awareness of initiating, executing, and controlling one's own volitional actions in the world.

- Normally SA and SO are tightly integrated, such that while e.g., walking one feels that "my own legs are doing the moving" (SO) and that "walking movements are volitionally controlled by me" (SA) [S. Gallagher, TICS, 2000]
  - In schizophrenia, the integration of SA and SO may become disrupted – i.e., actions may be executed, for which schizophrenic patients have a sense of ownership, but not a sense of agency, which could be attributed to God, ET, CIA, etc., which the patients feel are the agents who control the actions, while patients are aware that they execute the actions with their own bodies.

Alien hand syndrome (AHS): disorder of the SA

- AHS is a rare neurological disorder: the person may sometimes reach for objects and manipulate them without wanting to do so, even to the point of having to use the controllable hand to restrain the alien hand.

- Patients complain of a feeling of "strangeness" in relationship to the goal-directed movements of their hand and insist that "someone else" is moving the hand, and that they are not moving it themselves (SA is absent while SO is intact).
- AHS may happen when the two hemispheres are surgically separated, in some cases after brain surgery, stroke, infection, tumour, aneurysm and specific neurological degenerative disorders.

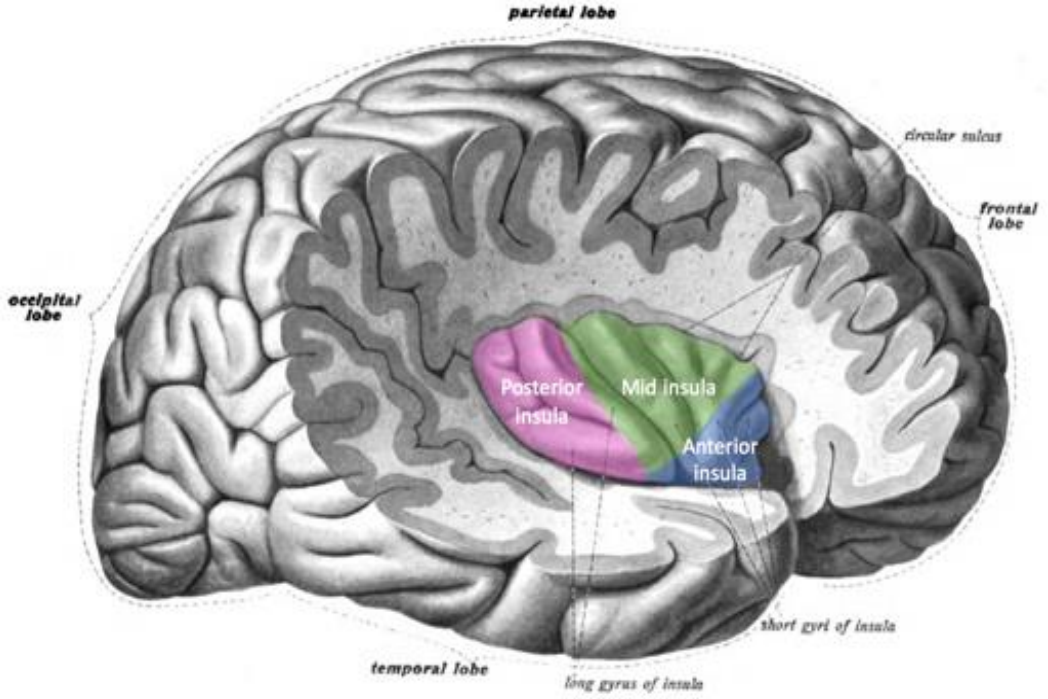
Disorders of the sense of ownership

- Neuroimaging experiments (PET, fMRI) have reported that activation of the junction between the **right** inferior parietal lobe and temporoparietal lobe (**temporoparietal junction, TPJ**) correlates with the subjective sense of ownership in action execution [Ruby and Decety, Nature, 2001; Farrer et al, 2003].

- In addition, lesions of this region can produce a variety of disorders associated with body knowledge and self-awareness such as
  - **Anosognosia**: a person who suffers some disability is unaware of the existence of his or her disability (paralysis).
  - **Asomatognosia**: a form of neglect in which patients deny ownership of their limbs.
  - **Somatoparaphrenia**: is a type of delusion where one denies ownership of an entire side of one's body.

Neuroscience of sense of agency

- Attributing an action to oneself activates the **anterior insula bilaterally** [Farrer and Frith, NeuroImage, 2002].
- Insula (or insulary cortex) is located in the TPJ. The anterior insula processes a person's sense of disgust both to smells and to the sight of contamination and mutilation — even when just imagining the experience. It is involved in primordial emotion or primordial feeling, which is an attention-demanding feeling (e.g., thirst, pain, fatigue) evoked by an internal body state.



Mirror neurons

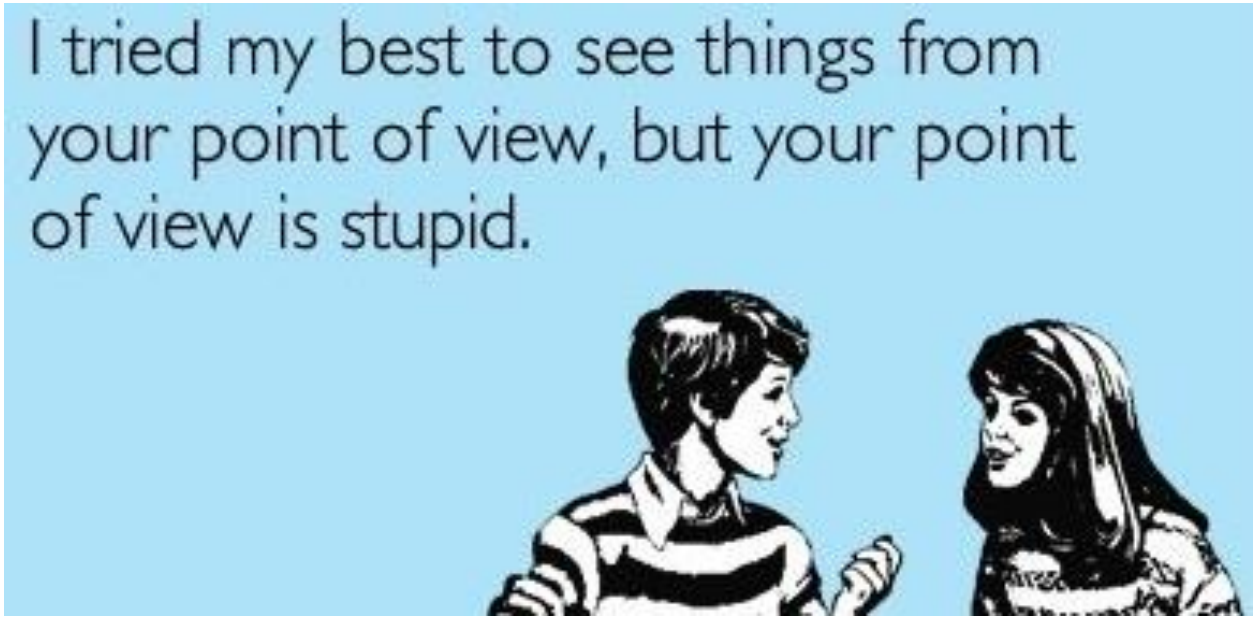
- Mirror neurons are neuron that fire both when an animal acts and when the animal observes the same action performed by someone else. Thus, the neurons "mirror" the behaviour of the other, as though the observer were itself acting.
  - Such neurons have been observed in primates (Rizzolatti et al. 1999), birds (Prather et al. 2008) and humans (Mukamel et al. 2010).

- In humans, brain activity consistent with that of mirror neurons has been found in the premotor cortex, the supplementary motor area, the primary somatosensory cortex, and the inferior parietal cortex.

- Some researchers speculate that mirror systems may simulate observed actions, and thus contribute to theory of mind skills. Neuroscientists such as Marco Iacoboni (UCLA) have argued that mirror neuron systems in the human brain help us understand the actions and intentions of other people.

Theory of mind (ToM)

- ToM is the **ability to attribute mental states—**beliefs, intentions, desires, knowledge, etc.—to oneself and others, and to understand that others have beliefs, desires, intentions, knowledge, etc. that are *different from one's own*. ToM is used to explain and predict the behaviour of others.
  - Deficits can occur in people with autism spectrum disorders, schizophrenia, attention deficit hyperactivity disorder, as well as alcoholics with the brain damage due to alcohol's neurotoxicity.



Why "theory of mind"?

- Theory of mind is called a theory. **Why the theory?**

- Because each human can only intuit the existence of their own mind through introspection, and no one has a direct access to the mind of another.
  - Intuition: The faculty of knowing or understanding something without reasoning or proof.

- It is typically assumed that other people have minds by analogy with one's own, and this assumption is based on the reciprocal, social interaction, as observed in joint attention, the functional use of language, and the understanding of others' emotions and actions (which in turn are based on own experience or imagery).

Empathy

- **Empathy** is the *capacity to feel emotionally* what another being (a human or animal) is experiencing from within the other being's frame of reference, i.e., the capacity to place oneself in another's position. It also is the ability to feel and share another person's emotions.
  - From *empatheia*, meaning "physical affection or passion" in Greek.

- ToM and empathy usually go hand in hand, but do not have to.

- Note that in English, sympathy (from the Greek words syn "together" and pathos "feeling" which means "fellow-feeling") is the perception, understanding, and reaction to the distress or need of another human being.

Development of ToM

- ToM appears to be an *innate potential ability* in humans: however it **requires social and individual experience** over many years of life for its full development.
  - Different people may develop more, or less, effective theories of mind, depending on their innate dispositions and particular experience.

- One of the most important milestones in theory of mind development is gaining the ability to attribute false belief: that is, to recognize that others can have beliefs about the world that are diverging.

- To do this, it is suggested, one must understand how knowledge is formed, that people's beliefs are based on their knowledge, that mental states can differ from reality, and that people's behaviour can be predicted by their mental states.

Sally-Anne false-belief task / test

- The tested child passes the task if she answers that Sally will look in the basket, where she put the ball.

- The child fails the task if she answers that Sally will look in the box, where the child knows the ball is hidden, even though Sally cannot know this, since she did not see it being hidden there.

Sally-Anne test: results

- To pass the task, the child must be able to understand that another's mental representation of the situation is *different* from their own, and the child must be able to *predict behaviour* based on that *understanding*.

- Most normally developing children are able to pass the task from around **age four**.

- However, 80% of children diagnosed with autism were unable to do so.

AUTISM

Persons with autism may possess the following characteristics in various combinations and in varying degrees of severity.

Inappropriate laughing or giggling

No real fear of dangers

Apparent insensitivity to pain

May not want cuddling

Sustained unusual or repetitive play; Uneven physical or verbal skills

May avoid eye contact

May prefer to be alone

Difficulty in expressing needs; May use gestures

Inappropriate attachments to objects

Insistence on sameness

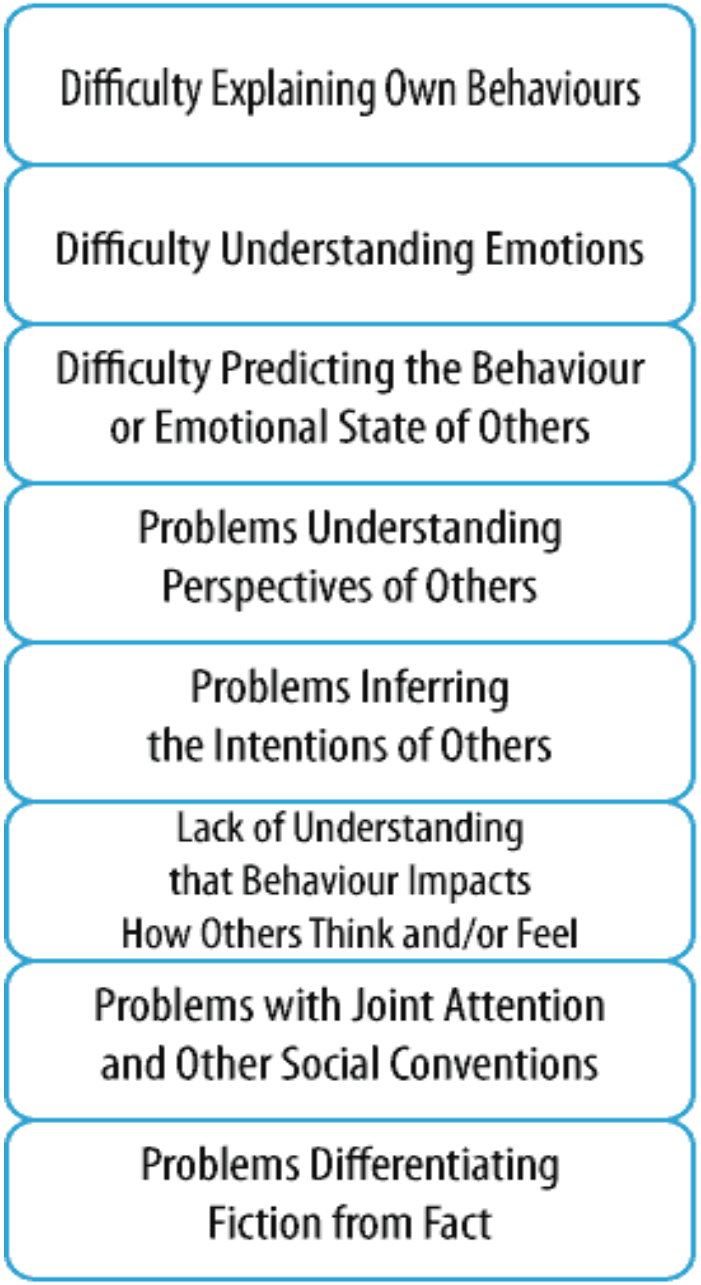
Echoes words or phrases

Inappropriate response or no response to sound

Spins objects or self

Difficulty in interacting with others

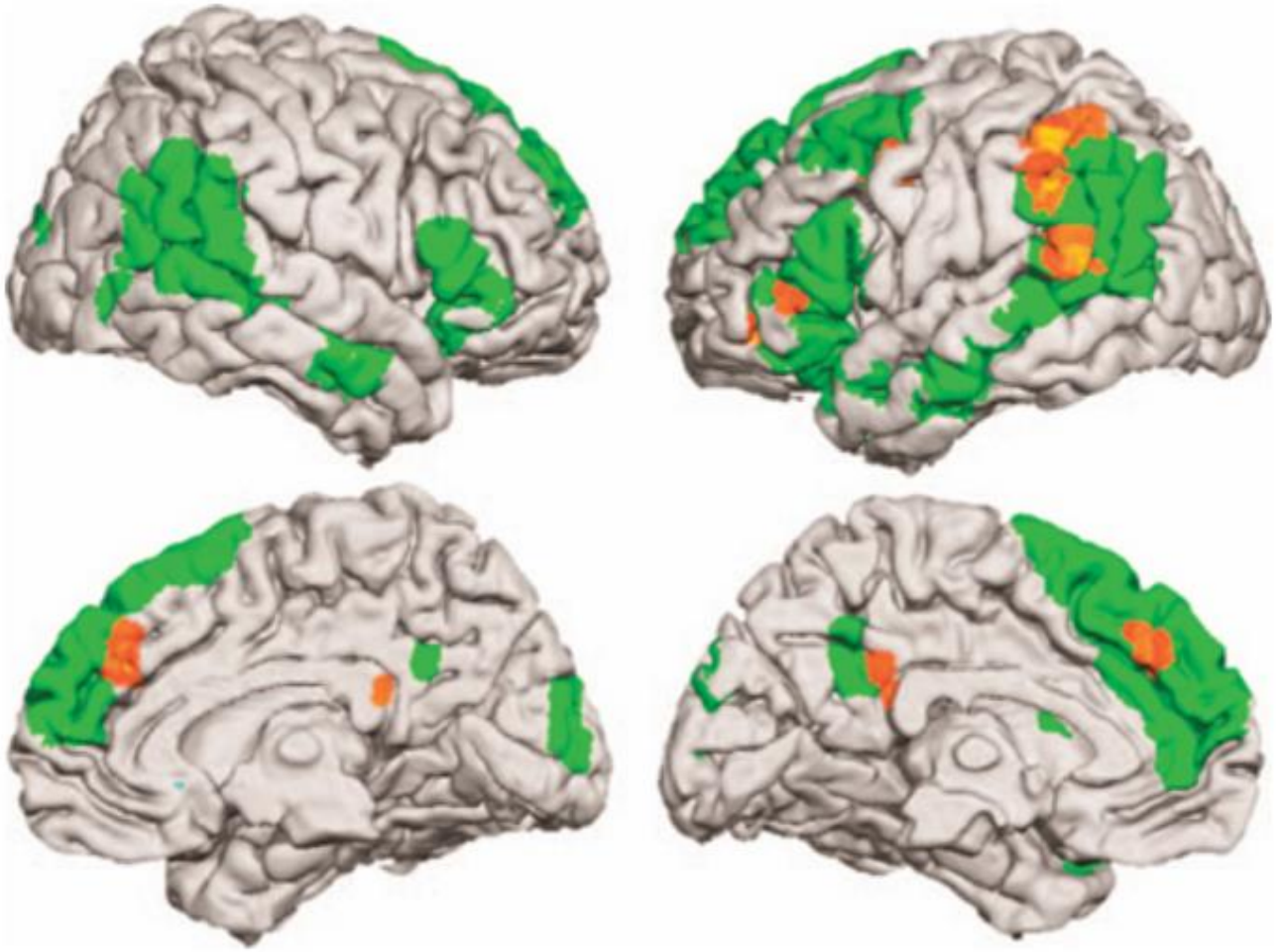
ToM and autism



- Autism is a neurodevelopmental disorder characterized by impaired social interaction, impaired verbal and non-verbal communication, and restricted and/or repetitive behaviour.
- There is no "one autism". It's a spectrum with varying degree of severity and mixture of symptoms.
- Baron-Cohen and Ute Fritz were first with hypothesis ToM is damaged in autism too.

Neuroscience of ToM in healthy subjects

- Cerebral activation (green), key areas of the ToM-Network (in red). ToM network encompasses the lateral, medial and orbital PFC (including areas with mirror neurons), lateral parietal and inferotemporal cortex and TPJ [Walter et al., Mol. Psychiatry, 2011].



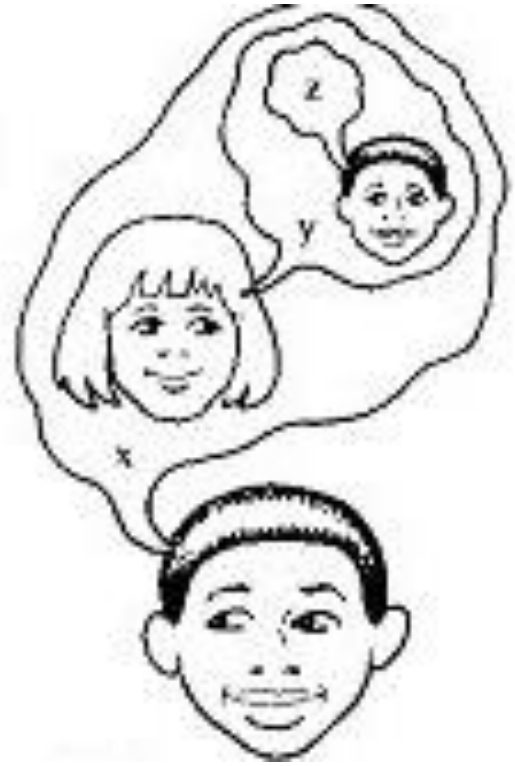
Consciousness: some basic terms

- **Arousal**: physiological and psychological state of being awake. *A necessary condition for awareness and consciousness.*

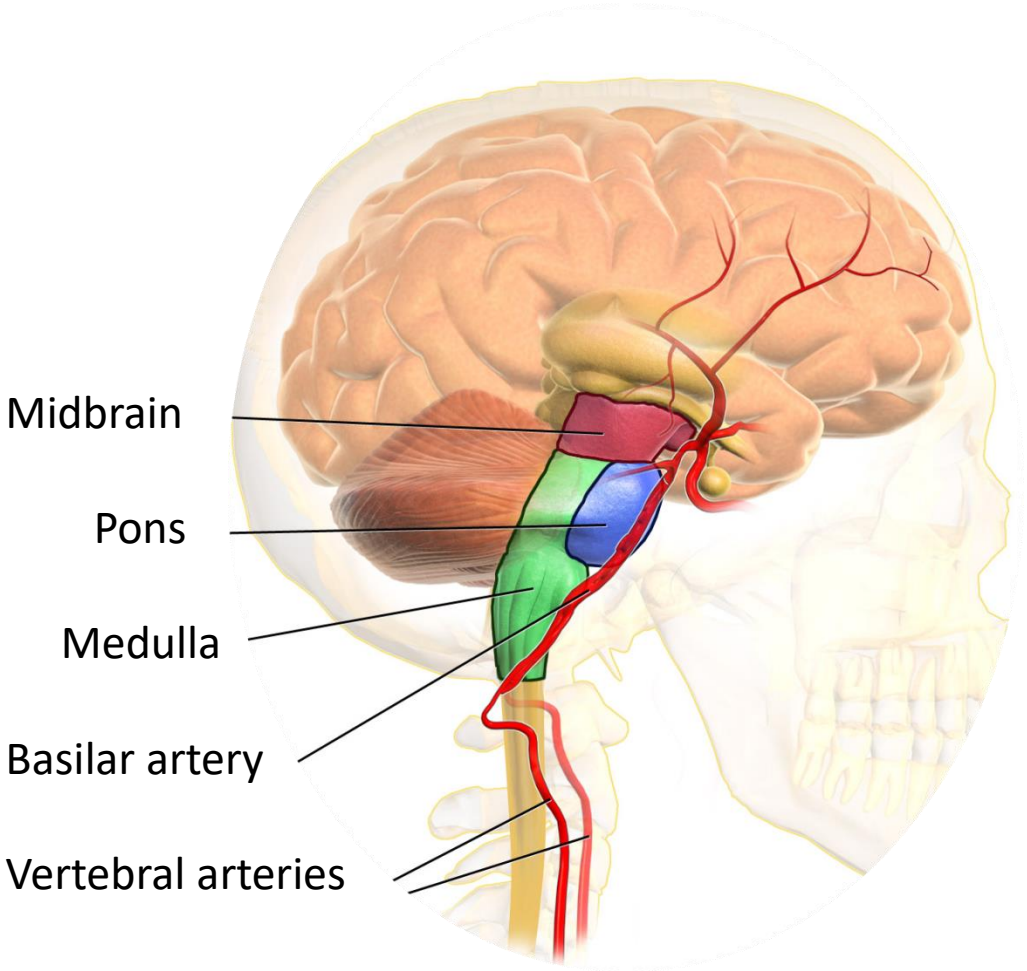
- **Awareness**: perceptual consciousness, being aware of something that is perceived.

- **Consciousness**: being aware of something within oneself, such as ideas, thoughts, emotions, feelings, self, etc., i.e. the **inner mental self-reflection**.

- **Altered consciousness**: psychoactive drugs, confusion, delirium, stupor, sleep, general anaesthesia, vegetative state, coma.



Arousal



- Arousal involves activation of the **reticular system in the brainstem** leading to sensory alertness, mobility and readiness to respond.

- Major systems responsible for arousal originating in the brainstem send connections throughout the cortex, and release neurotransmitters like **acetylcholine, norepinephrine, dopamine, histamine, and serotonin**.

Symptoms of **altered** consciousness (AC)

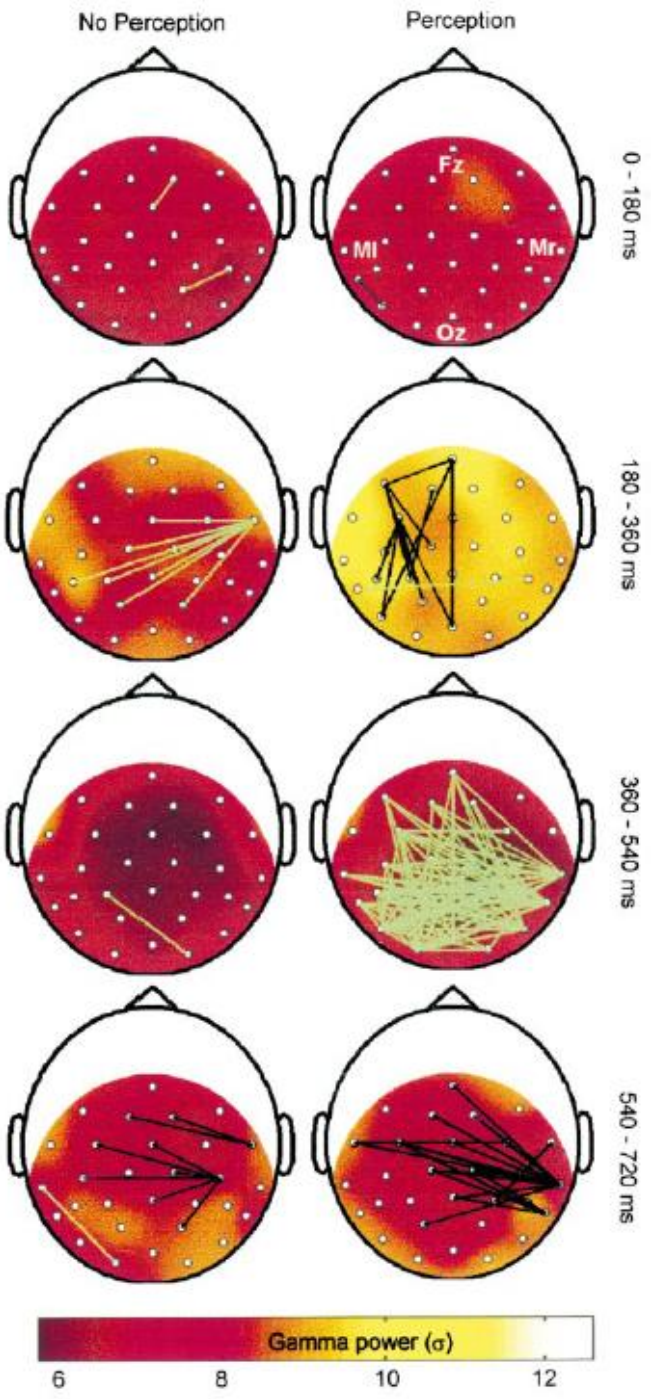
- Psychoactive drugs: alcohol, illegal drugs (hashish, meth, LSD, cocain, heroin, etc.). Effect depends on which neurotransmitters are affected.
- Sleep: naturally recurring and reversible state of mind characterized by altered consciousness, inhibited sensory activity, inhibition of voluntary muscles, and no interactions with surroundings.
- General anaesthesia: unconsciousness, amnesia, analgesia, loss of reflexes of the autonomic nervous system, resulting from the administration of one or more general anaesthetic drugs.
- Persistent vegetative state: the patient has sleep-wake cycles, but lacks awareness and only displays reflexive and non-purposeful behaviour.
- Coma: the patient lacks awareness and sleep-wake cycles and only displays reflexive behaviour.
- Brain death: The patient lacks awareness, sleep-wake cycles, and brain-mediated reflexive behaviour.

(Perceptual) Awareness

- In a pioneering study of Rodriguez et al. (Nature, 1999) humans were presented with the so-called Mooney faces either in the canonic upright position (**a**) or in an upside position (**b**).

- It is easy to recognize a human face in the canonic position (a) but very difficult to recognize a face when they are presented upside down (b).

- Viewing these images, either in a normal or turned position, is always accompanied with the increase of gamma activity in the visual areas.

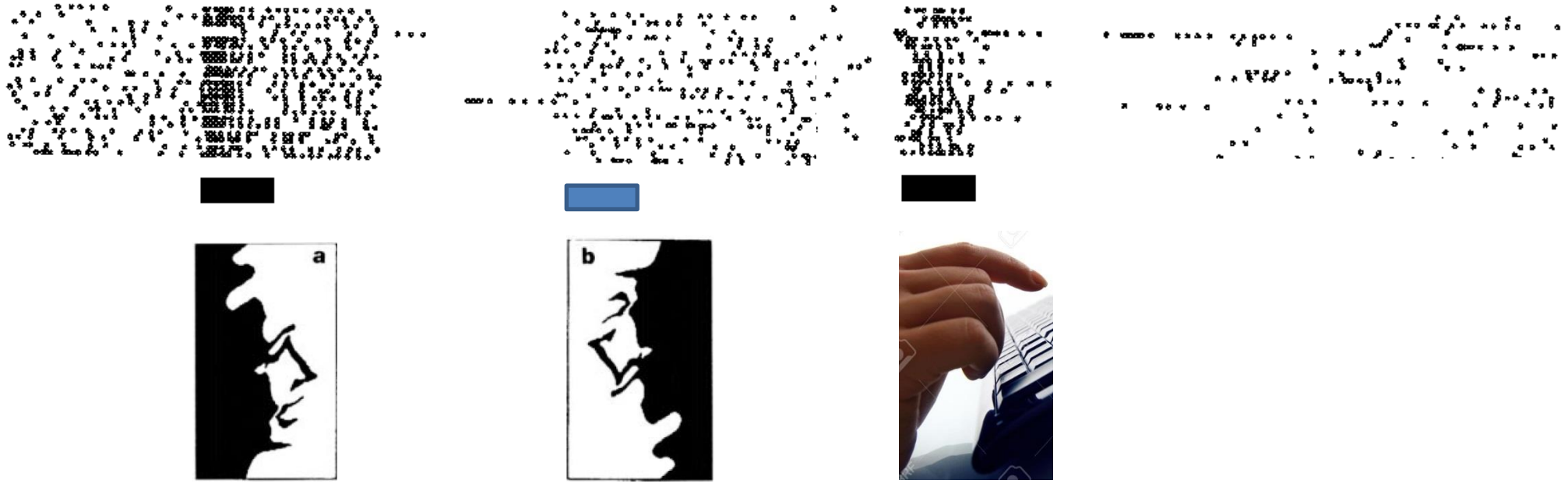


Awareness

- Precise transient synchronization of gamma oscillations occurs in the time window 180-360 ms after presentation of the stimulus, but only when the subject perceives a face.
- It is intriguing that this synchronization occurs only in the left hemisphere which is the so-called conscious hemisphere.
- When the subject did not perceive a face, but instead only a nonsense black-and-white patches, no synchronization happened in the cortex.
- The 2nd transient synchronization occurred in the premotor and motor areas of both hemispheres during the motor response of subjects (540-720 ms)

Illustration of coherent and asynchronous activity

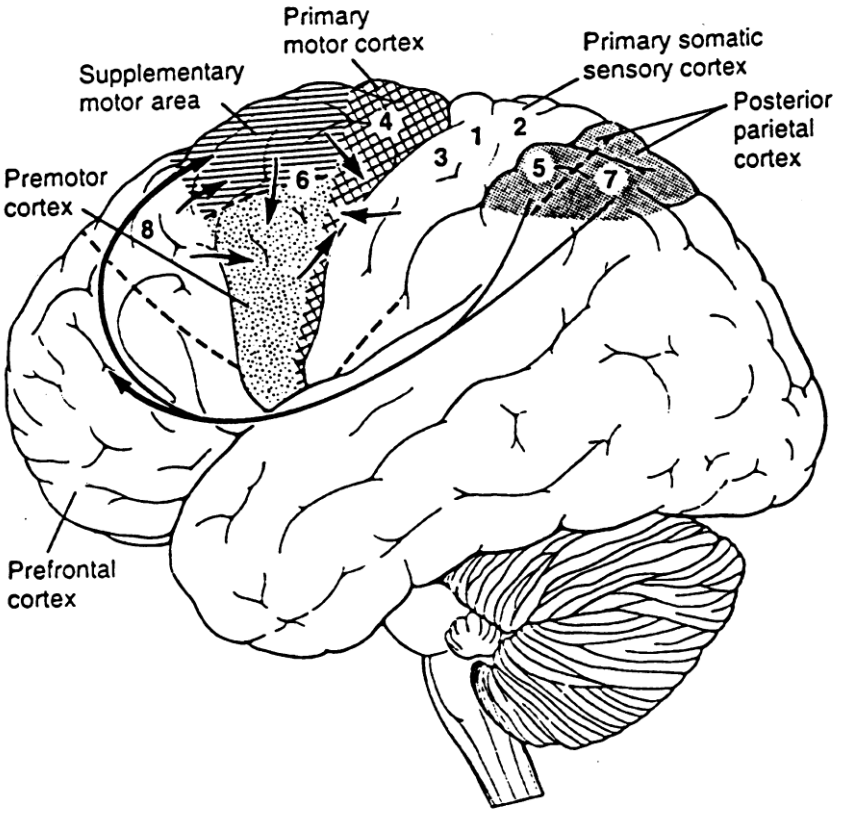
- Illustration of the neural activity during synchronization and asynchronous activity (dots are spikes of individual neurons).



- When neurons fire simultaneously within a narrow time window then we say they have a coherent / synchronous activity. Asynchronous spiking means absence of coherence between neurons.

Awareness

- Rodriguez et al. (1999) has demonstrated that **perception** of faces in humans is **accompanied by a transient (~180 ms) synchronization of gamma activity** in hierarchically highest visual areas in the parietal cortex and premotor areas in the frontal cortex (Kandel et al. 1991).

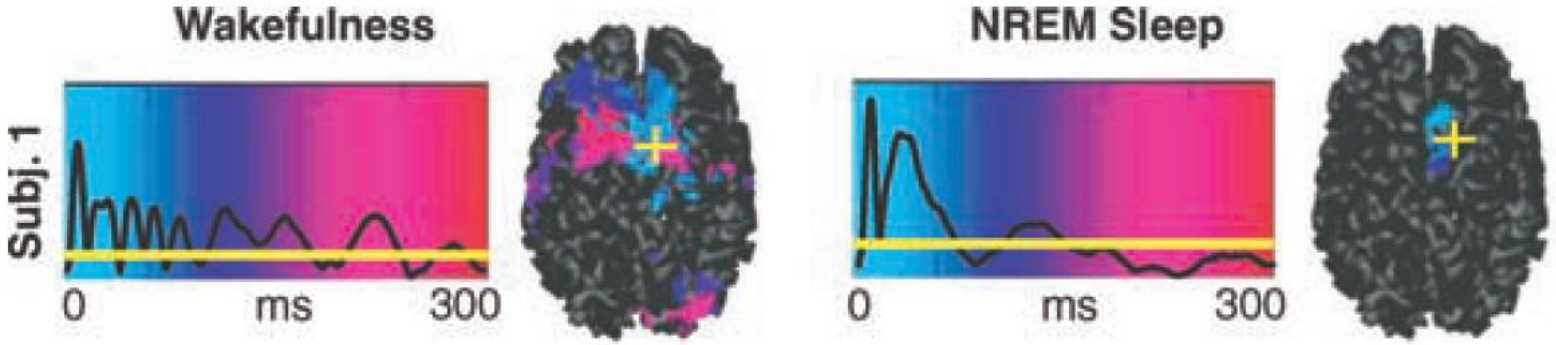


- Thus, realization that we see some object that we can identify (i.e., a face) takes time, it is not instant. And this time is cca 180 ms.

Breakdown of cortical effective connectivity during sleep

- When we fall asleep, consciousness fades yet the brain remains active.

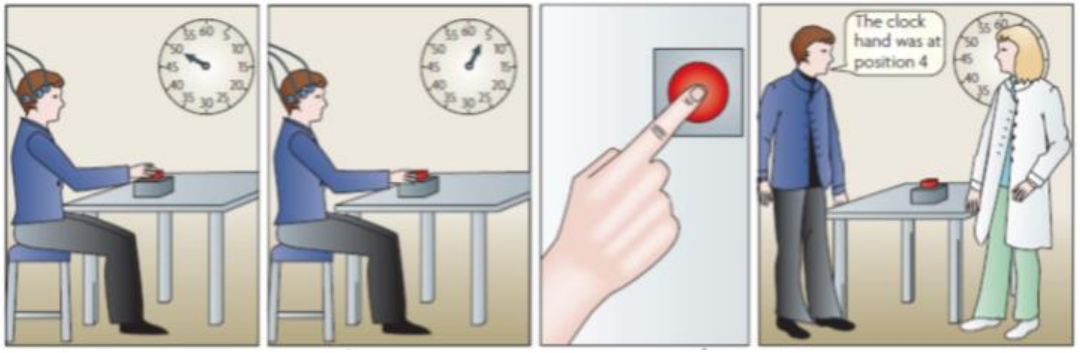
- Massimini et al. 2005 used TMS together with HD-EEG and asked how the activation of one cortical area (the premotor area) is transmitted to the rest of the brain. During quiet wakefulness, an initial response at the stimulation site was followed by a sequence of waves that moved to connected cortical areas several cm away.
- During non-REM sleep, the initial response was rapidly extinguished and did not propagate beyond the stimulation site.



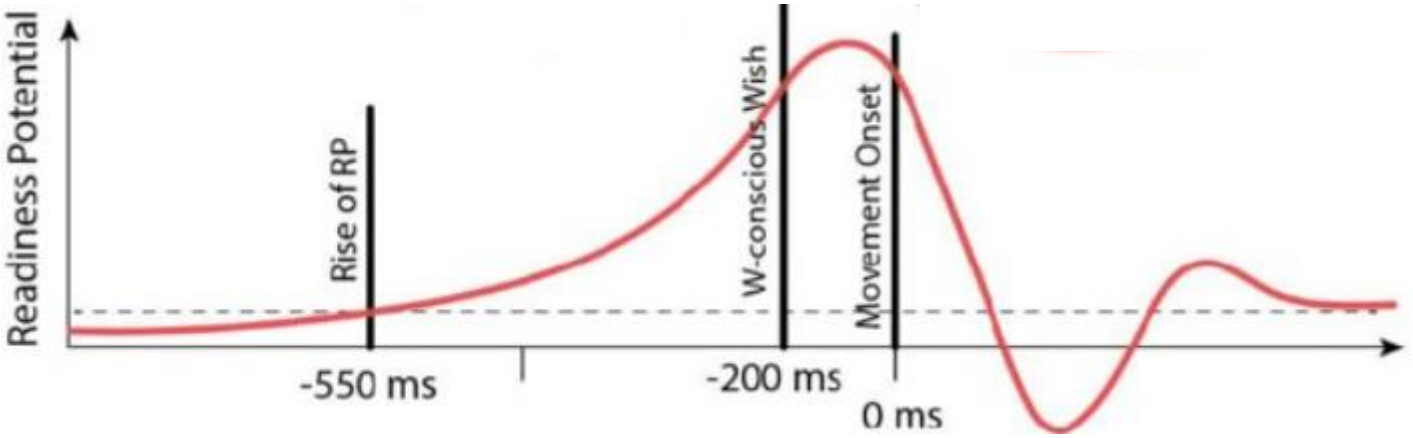
- Thus, the fading of consciousness during certain stages of sleep may be related to a breakdown in cortical effective connectivity.

Free will and the brain

- In an experimental study (Libet et al. 1983), subjects pressed the key whenever they "felt the urge" to do so (free will). At the same time, participants had to watch a clock-like counter to report the exact time they felt the urge to move.



- The onset of the brain preparatory activity (readiness potential) preceded the appearance of the subject's awareness of the conscious wish to act, by about 350 msec. That indicated that the volitional process is initiated unconsciously.



Free will and the brain

- Subjects have no reportable awareness or intuitive feeling that the brain has started a process of planning and preparation before their conscious wish/urge to act appears.

- Unconscious initiation of the voluntary process means that conscious free will could not actually 'tell' the brain to begin its preparation to carry out a voluntary act !

- However, the conscious free will could block or 'veto' the process, resulting in no motor act (Libet, 1985). That presumably occurs when a given wish is recognized as being incompatible with social acceptability and with one's personality.

- However, some are sceptical because the question remains how the mental process affects electrical activity of the brain? And does it?

Consciousness of artificial systems

- **Turing test** (1950): To pass the test, an AI system must be able to imitate a human well enough to fool interrogators. The Turing test is commonly cited in discussions of AI as a proposed criterion for machine consciousness; it has provoked a great deal of debate.

- Daniel Dennett and Douglas Hofstadter argue that anything capable of passing the Turing test is necessarily conscious, while **David Chalmers** argues that a philosophical zombie could pass the test, yet fail to be conscious.

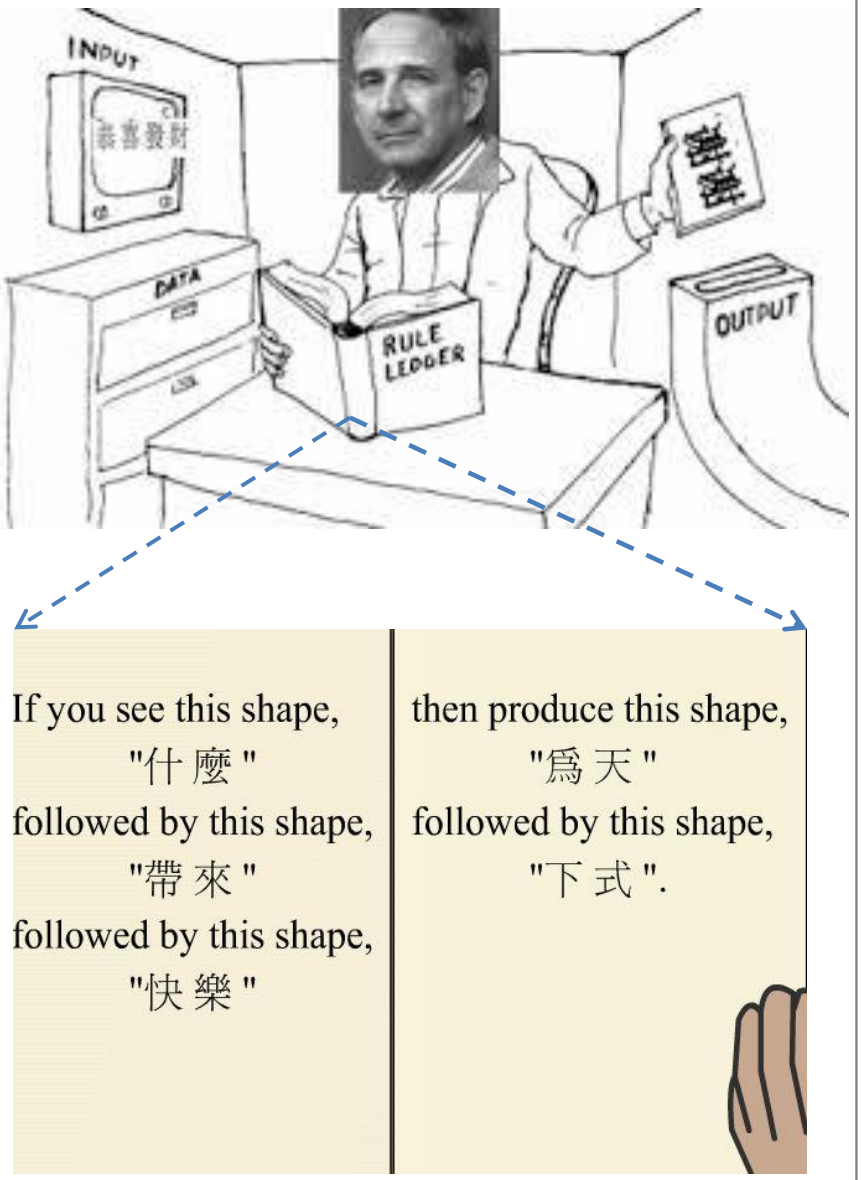
- A philosophical zombie or p-zombie in the philosophy is a hypothetical being that is indistinguishable from a normal human being except in that it lacks conscious experience, qualia, or sentience.
  - E.g, a p-zombie could be poked with a sharp object, and not feel any pain sensation, but yet, behave exactly as if it does feel pain (i.e., say "ouch" and recoil from the stimulus).

The Chinese room argument

- John Searle refuses the claim of proponents of what he calls "strong AI" that a computer program can be conscious, though he does agree with advocates of "weak AI" that computer programs can be taylored to "simulate" conscious states.

- Searle proposed that "*causal powers*" of the sort that the brain has and which the AI systems lack, are needed for consciousness.

- I.e., conscious person can perform computations, but consciousness is not inherently computational the way computer programs are.



Neural correlates of consciousness

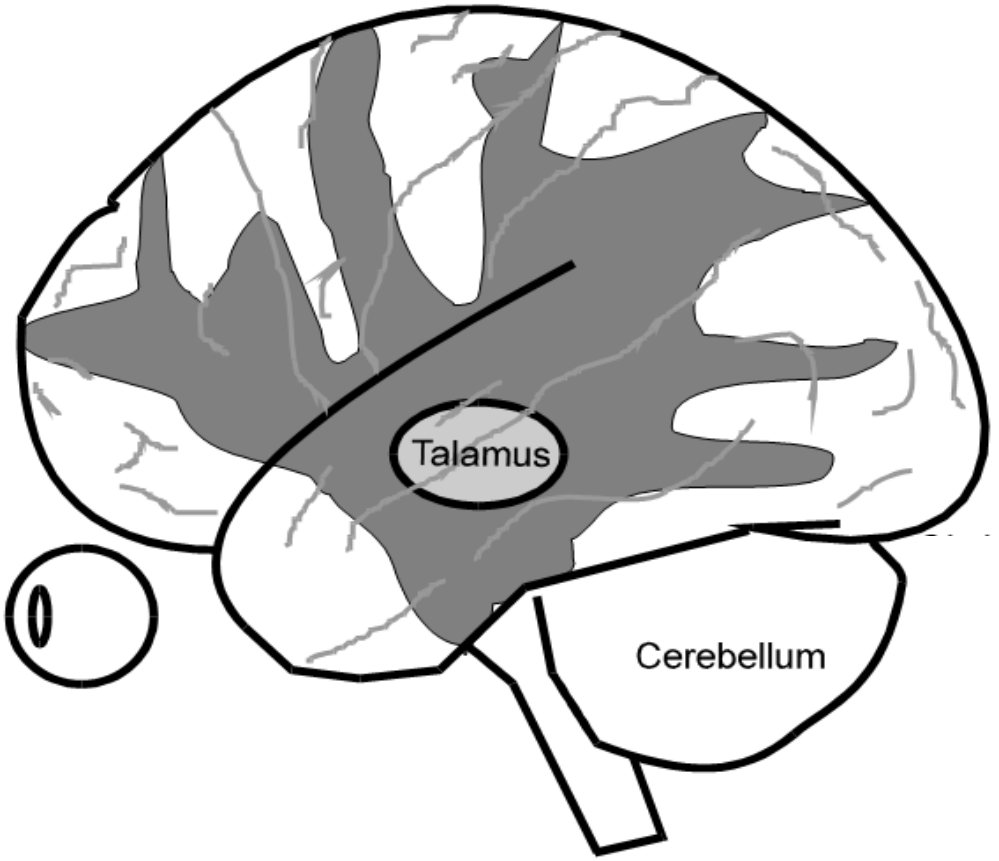
- Based on analysis of neurobiological experiments, F. Crick, C. Koch, W. Singer, G. Edelman and G. Tononi advocate essentially this hypothesis of neural correlate of consciousness:
  - *Consciousness is accompanied by the fast semiglobal coherent activity of the brain, called the dynamic core.*

- The dynamic core corresponds to a large (semiglobal) continuous cluster of neuronal groups that are temporarily coherently active on a time scale of hundred of milliseconds (ms).

- Neuronal groups participating in the DC are much more strongly interactive among themselves than with the rest of the brain.
  - However, the dynamic core must also have an extremely high complexity as opposed to for instance epileptic convulsions, which are dynamically quite simple.

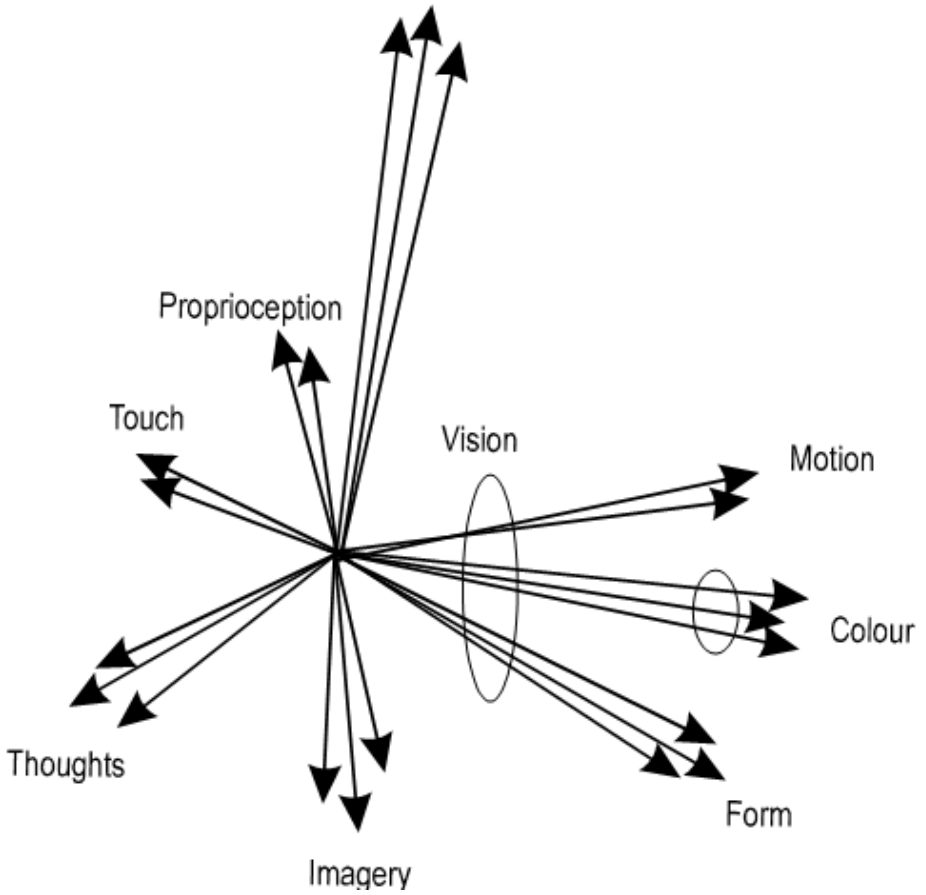
The dynamic core (is it "Searle's causal power"?)

- Illustration of the dynamic core, a **changing coherent semiglobal activity** of the brain.
- The dynamic core consists of a large number of distributed groups of neurons which enter the core temporarily based on their mutual coherence over the **time span of cca 150ms**.



Interpretation of the dynamic core

- Neural reference space for any conscious state may be viewed as an abstract **N-dimensional neural reference space**, where each axis (i.e. dimension) stands for some participating group of neurons that represent a given aspect of the conscious experience (~ 10^{4-5} dimensions).



Emergence of the dynamic core

- Connecting different groups of neurons into temporarily synchronized whole requires dense **recurrent connections** between brain areas across all the lobes, thalamus and other subcortical areas, along which a **reiterated re-entry** of neural signals occurs.

- Each roughly 150 ms, a pattern of semiglobal activity must be selected within less than a second out of a very large, almost infinite, repertoire of options.
  - Thus, the dynamic core changes in composition over time.

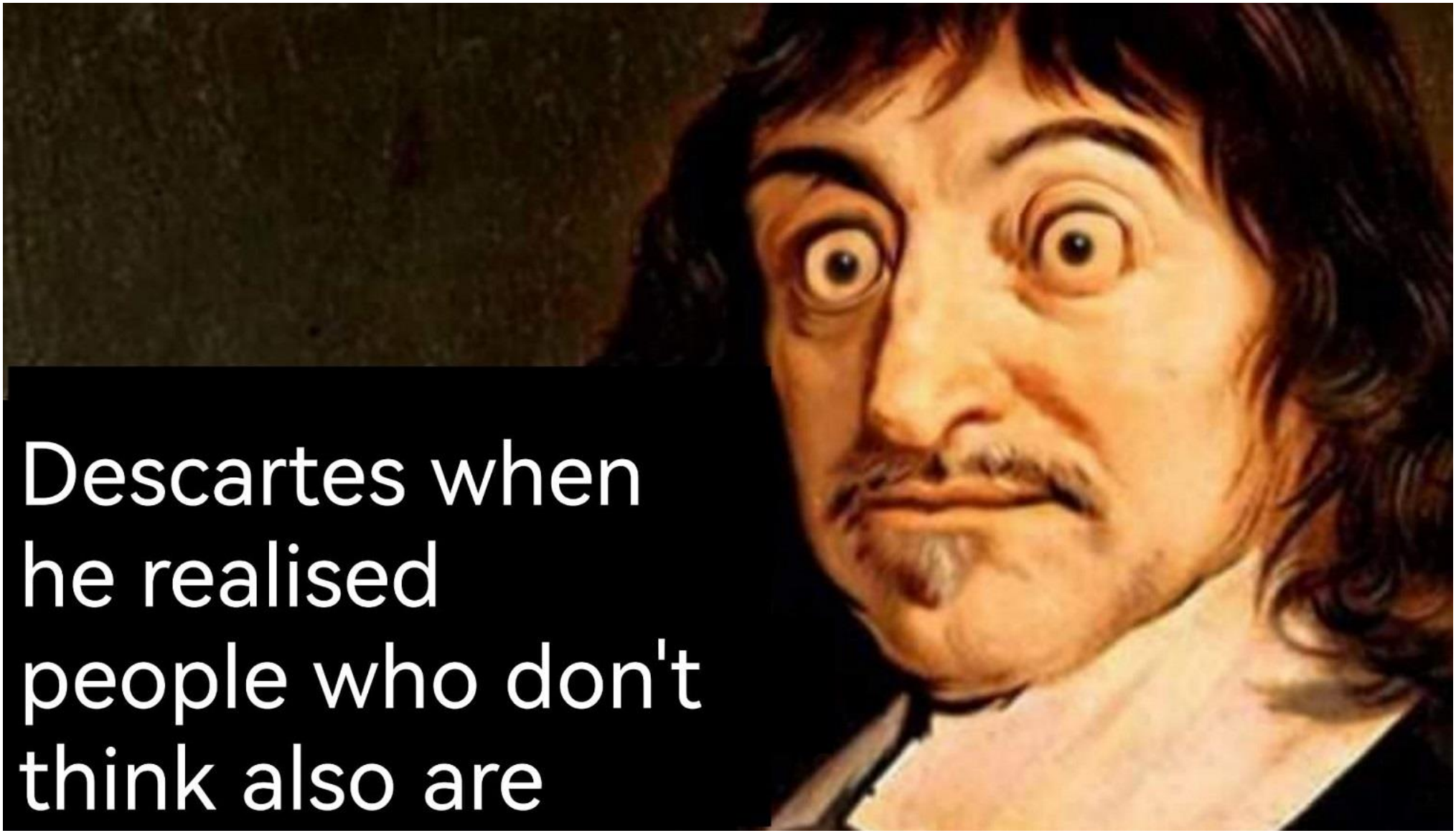
- **As suggested by neuroimaging, exact composition of the core related to particular conscious states vary significantly not only over time within one individual, but also vary significantly across individuals for the same stimulus (**is this the basis of qualia**?).**

Consciousness and quantum mechanics

- The constitutive elements of consciousness are **qualia**. The *subjective feelings* associated with the **redness of red** or the ***painfulness of a toothache*** are two distinct qualia. So far, it remains mysterious how the physical world gives rise to such sensations (the hard problem of consciousness according to Chalmers.)
- Physicist **Roger Penrose** has claimed that brains can evaluate noncomputable functions; that this ability is related to consciousness; that both this ability and consciousness require a yet-to-be-discovered *theory of quantum gravity* and that microtubules are the sites of the associated quantum gates (i.e. basic quantum circuits).
- **Koch & Hepp** (Nature, 2006) massively **criticise this idea** on a number of grounds. E.g., even if quantum gates exist inside the trillions of neurons, it remains totally nebulous how information of relevance to the organism would get to these quantum gates.

The discourse goes on…

- The two traditional and competing theories of consciousness are **dualism and materialism**.
- While there are many versions of each, the dualism generally holds that the consciousness is non-physical in some sense, whereas the materialism holds that mind / consciousness **is** the brain.
- *Objection to materialism*: For example, it is often said that materialism cannot truly explain how or why some brain states are conscious, and some are not; thus there is an important "explanatory gap" between mind and matter.
- *Objection to dualism*: how a non-physical substance or mental state can causally interact with the physical brain?
- **Or is the mental state / consciousness some kind of yet-to-be-discovered physical quality existing in this physical Universe ???**


Descartes when
he realised
people who don't
think also are