# Integrated Information Theory of Consciousness

Neil Bramley

# Plan I

Neil (~25 minutes):

1. Background
    1. The hard problem of consciousness
    2. Functionalism > Computationalism
2. Integrated information theory 1.0
    1. Measuring information
    2. Measuring integrated information Φ
    3. Φ as *quantity* of consciousness
    4. Q-shapes as *quality* of consciousness
3. Some soft neurobiological support for IIT

# Plan II

George (~25 mintues):

1. Problems with computationalism in general
    1. Scope
    2. Liberalism/ panpsychism
    3. Spatiotemporal grain
2. Criticisms of IIT 1.0 specifically
    1. Silent units
    2. Nested consciousness and the exclusion principle
    3. Intractability
3. Responses by Tononi and Koch
4. Updated theory IIT 3.0
    1. Closer link with causal structure of operations system
    2. Demonstrable 'zombie' systems
    3. Charge of *ad hoc* -ness

# Plan III

After the break:

- 1-2 groups, open discussion of Integrated Information Theory, report back
- + Can also discuss related topics: e.g. functionalism, reduction, hard problem, mental causation

# The hard problem

- Chalmers (1995): We can think of the problem of consciousness into lots of 'easy' problems and a remaining 'hard' problem
- 'Easy' problems - How do we:
  - React to stimuli
  - Focus attention
  - Categorise/discriminate
  - Control behaviour
- 'Hard' problem –
  - Why are any of these processes accompanied by *experience?*
  - Why is there "something it is like" (Nagel, 1974) for us to discriminate the colour of a tomato?
  - Why not just the processing without the phenomenology?
- 'Explanatory gap' (Levine, 1983) between e.g. neural and psychological explanations for behaviour.
- Why is the one accompanied/paralleled by the other?

# Explananda of the theory

- Sentience
- Wakefulness
- Awareness
- Access consciousness
- Self-consciousness
- Narrative consciousness

Extra baggage

- Experience/subjectivity
  - Phenomenology
  - Qualia
  - "*What it is like?*" (Nagel, 1974)

Core concept

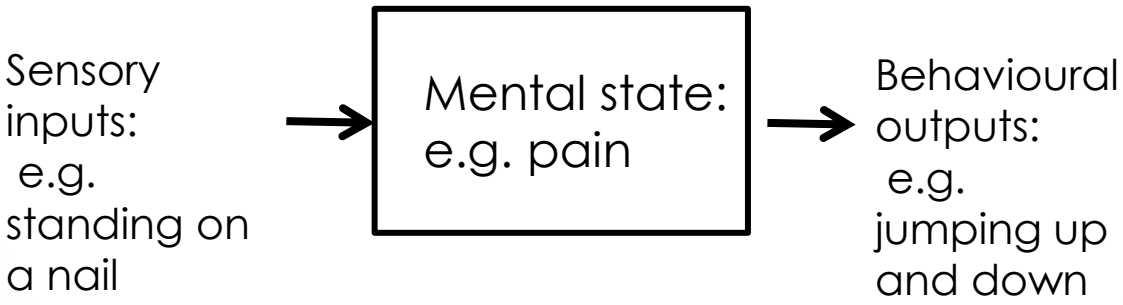
# Multiple realizability -> Early functionalism

- Traditional psychophysical identify theories (e.g. Place 1956; Smart, 1959) identified mental states directly with physical states

pain = c-fibers firing

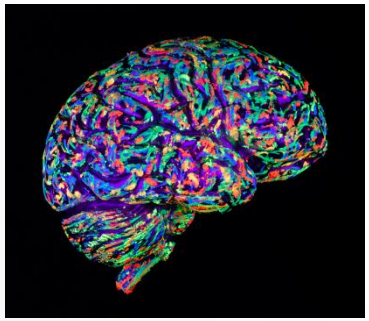
- But Putnam(1967) argues that for any mental to physical correlation, one can conceive of a different physical substrate giving rise to that same mental state
- If so, mental states are multiply realizable + identity between mental states and physical states is *contingent* rather than *necessary*
- Traditional functionalist theories instead identify mental states with their *functional role*.
- I.e. It is what the mental state *does* that matters, not what it is *made of*

# Functionalism



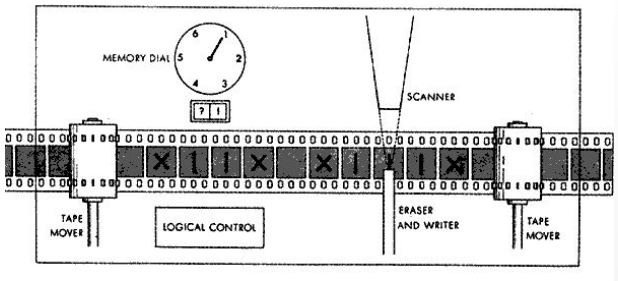
# Functionalism->Behaviourism

- Early functionalism movement became associated with the behaviourist movement in psychology plus logical positivism in philosophy
- Became a process of reconceptualising mental states functional input-output predispositions:
  - E.g. "Henry has a toothache" becomes "Henry is disposed (all things being equal) to cry out or moan and to rub his jaw"
- Or in case of 'machine-functionalism', associating mental states with the machine tables of an hypothesised infinite Universal Turing Machine (Block, 1980)



Brain

==

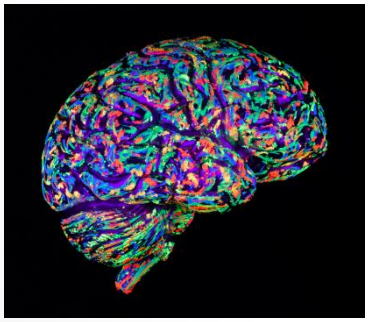



Universal Turing machine

# But functional identity at what level?

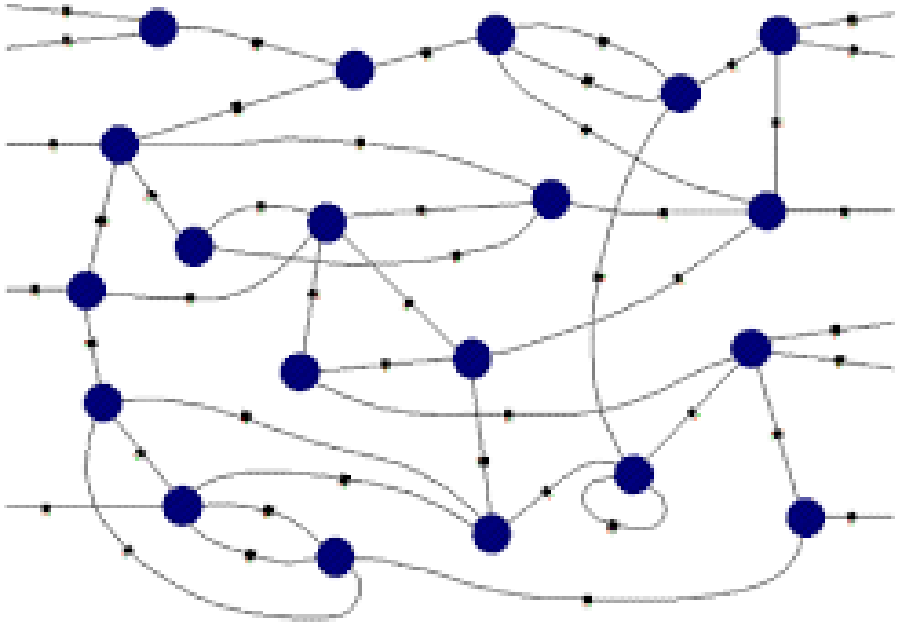
- With cognitive revolution, functional analysis of mental states in terms of inputs and outputs fell out of favour
- But understanding of computational systems also became more sophisticated:
- Marr (1972) defined three distinct levels of analysis for any system:
  - **The computational level** – What is the purpose of the system? What inputs should be mapped to what outputs?
  - **The process level** – How does the system achieve this? What representations does it use, and what processes does it employ to build and manipulate the representations?
  - **Implementational level** – How is the system physically realised? Is it built of neurons or silicon? How are they arranged?

- Early formulations of functionalism suggested claim of identity at the *computational* level
- Early physicalist identity claims at the *implementational* level.
- But very different *processes* (operations + representational structures inside a system) can achieve the same computational input-output relations
- Perhaps more fruitful to identify phenomenology with processing itself



Brain

Analogue
Parallel



# Functionalism -> Computationalism

- Resurgence of functionalist ideas about mental states along with "cognitive revolution"
- Computationalism (e.g. Edelman, 2008; Tononi, 2008):
  - The brain/mind is an information processing system
  - Thinking is a form of information processing
  - Phenomenology closely linked to cognitive processing
    - e.g. dynamics of information flow, representational structure
    - NOT to mapping of inputs to outputs
    - NOT to implementational substrate

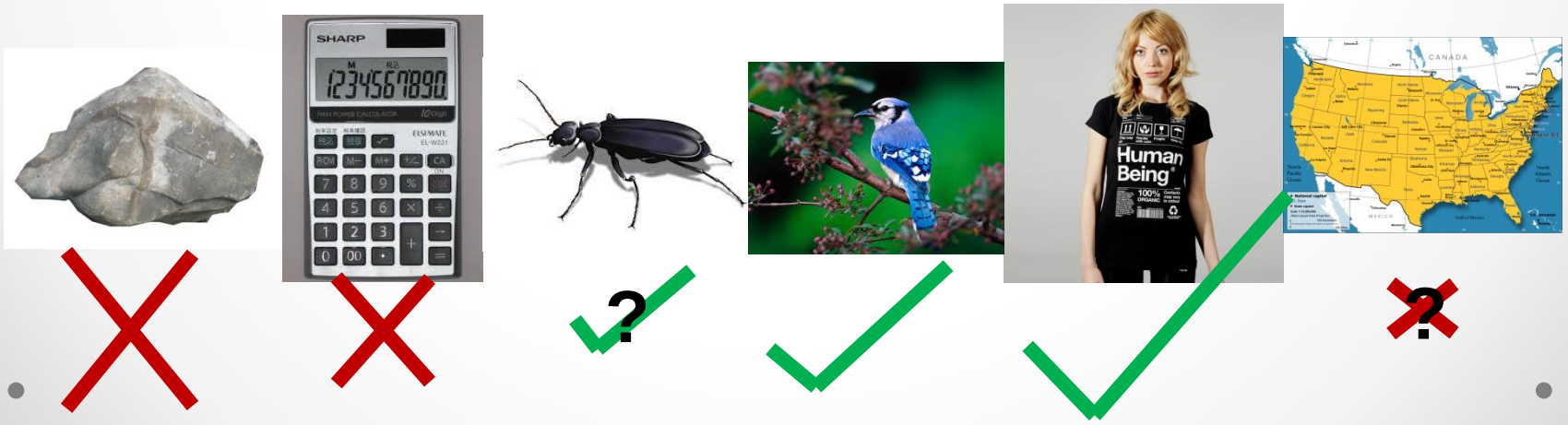
- Computationalism implicitly assumed in cognitive science. i.e. If the brain is not a computer, then why try to model it/ theorise about how it works?

# Integrated information theory

- Gilulio Tononi, Italian psychiatrist and neuroscientist develops a positive computational theory of consciousness, called *Integrated Information Theory* (IIT). References: Tononi, 2008; Edelman & Tononi, 2013; Balduzzi & Tononi, 2008; 2009; Oizumi, Albantakis & Tononi, 2014.
- IIT provides criteria for assessing the *quantity* and *quality* of consciousness present in any system

To get off the ground:

- Must make sensible predictions for systems that we have strong intuitions about



# Integrated information theory

**Claim: Consciousness *is* integrated information**

- **Information** – reduction in uncertainty
- **Uncertainty** – formalised as information entropy (Shannon, 1948)
- **Integrated information** ($\Phi$) – Information generated by a system over and above information generated by its parts taken separately

- **Quale/Qualia** – Specific individual instances of subjective conscious experience
- **Qualia space (Q)** – a multidimensional space, defining all possible states of the system, within which, any particular shape specifies a unique quale.

- **Claim (II) – The *quantity* of consciousness in a system is its $\Phi$ and the *quality* of consciousness is its shape in Q**

# What is information?

- Claude Shannon (1948) invented information theory to quantify the noise/uncertainty in communicating information down a phone line.
- Intuitively related to modern day concepts of 'bandwidth' and computer memory.
- Uncertainty H(x) is a measure of how much you don't know yet about x.

$$H(x) = -\sum_{i \in x} p(x_i) \log p(x_i)$$

- Can be measured in 'bits' (if we use log base 2).
- Complete uncertainty across $2^N$ possibilities = N bits
- Complete certainty = 0
- Information is quantified as *reduction in uncertainty (e.g.* given evidence)
- E.g. Toss a coin. Before you look at it you have 1 bit of uncertainty. By looking at it, you reduce space of possibilities from 2 to 1. You gain 1 bit of information.

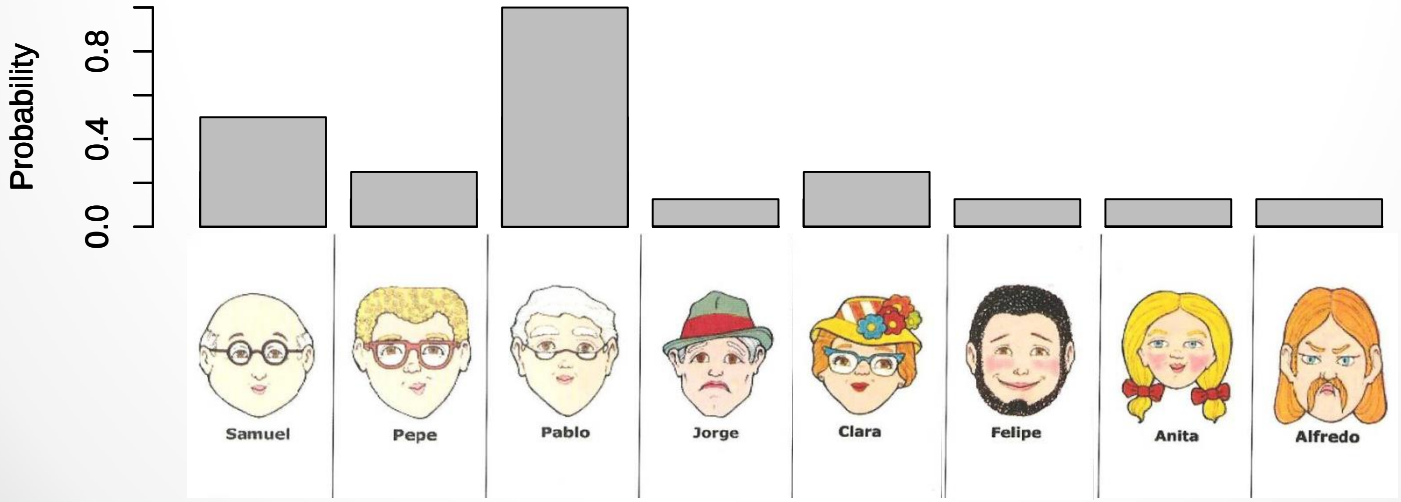
# Information example

- Are they wearing glasses?
- ✓ Yes
- Have they got white hair?
- ✓ Yes
- Are they bald?
- ❖ No

Uncertainy

$$H(x) = -\sum_{i \in x} p(x_i) \log p(x_i)$$

$$H(x) = -\left(\frac{1}{8} \log_2 \frac{1}{8}\right) \times 8 = 3 \text{ bits}$$

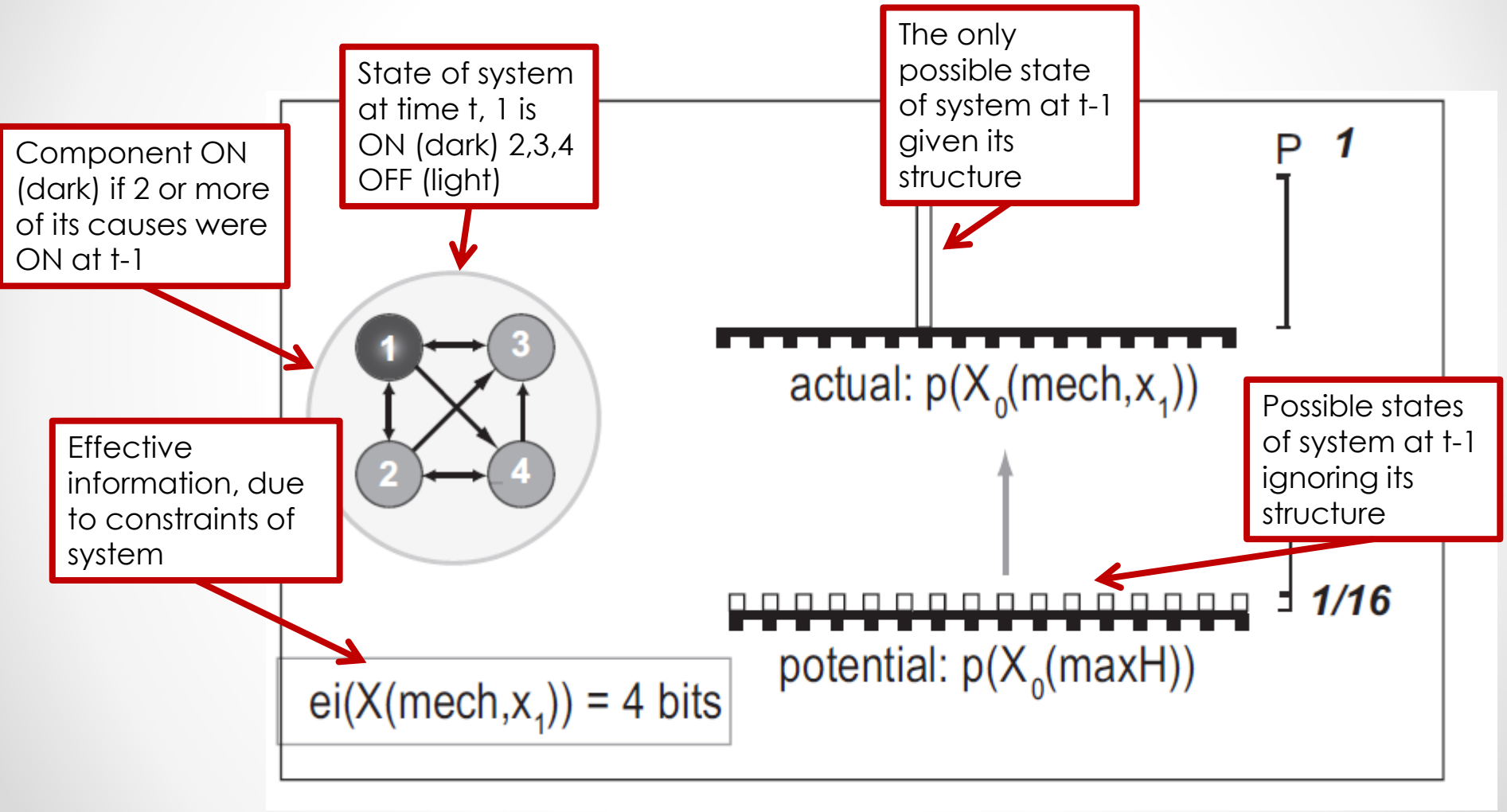


# Integrated information

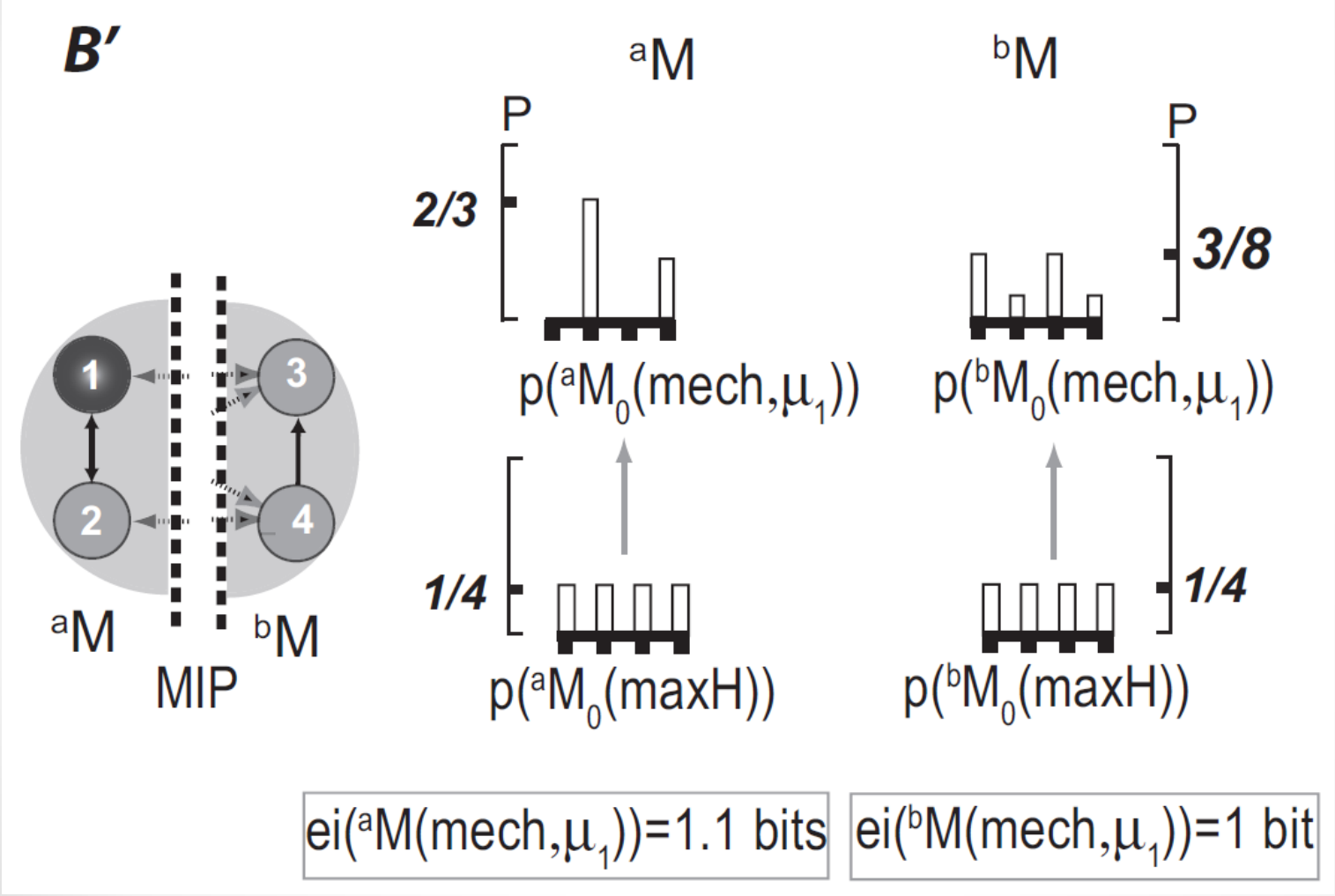
- Many systems involve lots of information but do not seem conscious:
- E.g. Digital camera, each pixel responds to the light. Taking a photo stores a lot of information
- From 'external perspective', we can look at a photo and see e.g. Marilyn Munro
- But: From 'internal perspective' of the camera, each pixel is independent and unconnected, just millions of light/dark discriminations.

- A richer system would *integrate* the information from the different pixels and use it to make broader range of discriminations than just light/dark, e.g. 'contains a face?', 'male or female?', etc.
- How can we capture the degree to which a system integrates information?

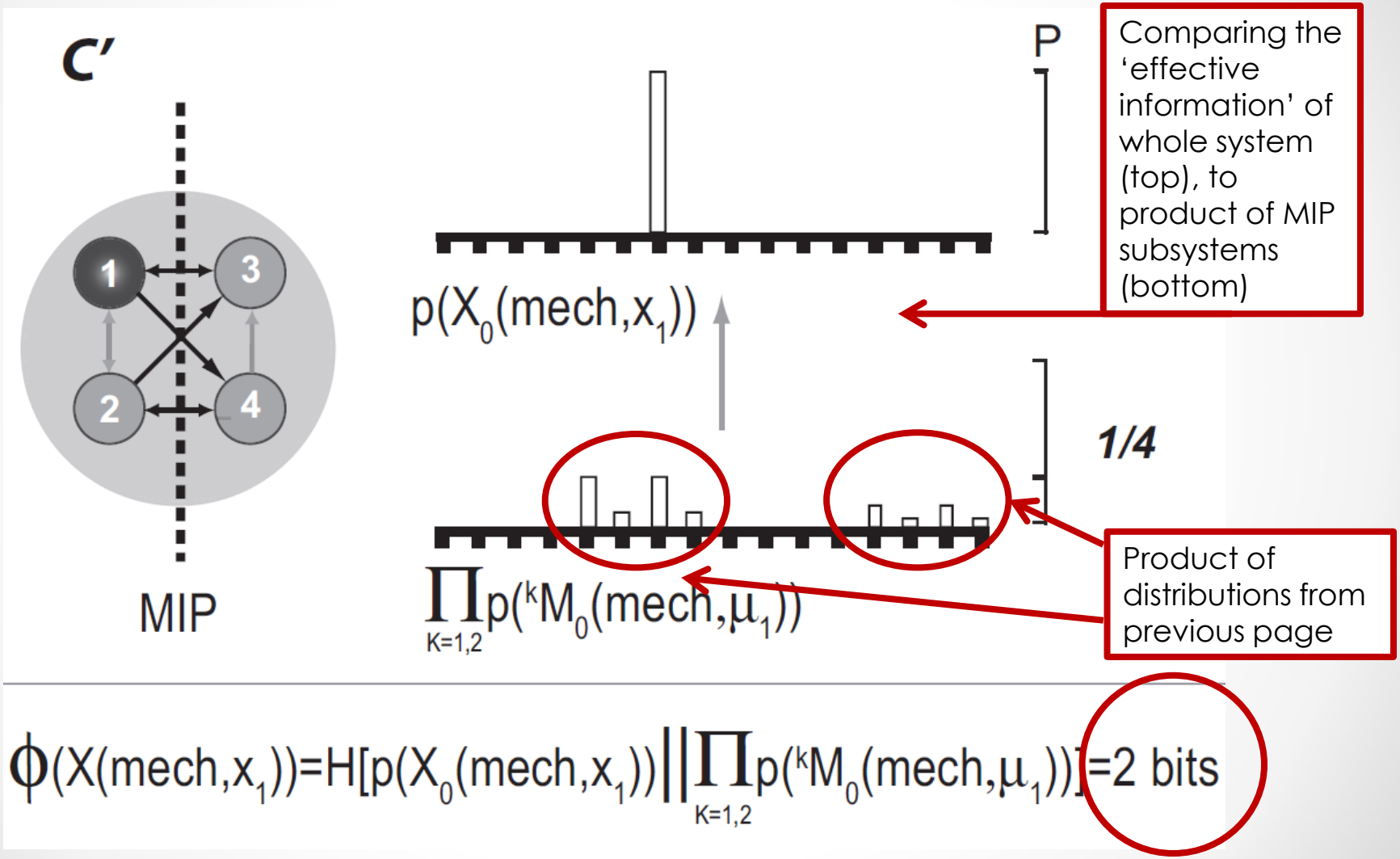
# Example system – Effective information



# Example system - Irreducible information



# Example system

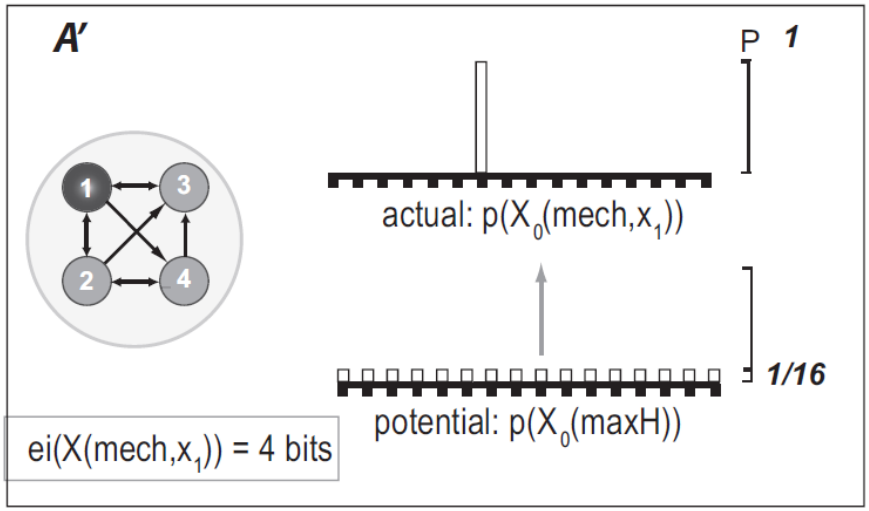


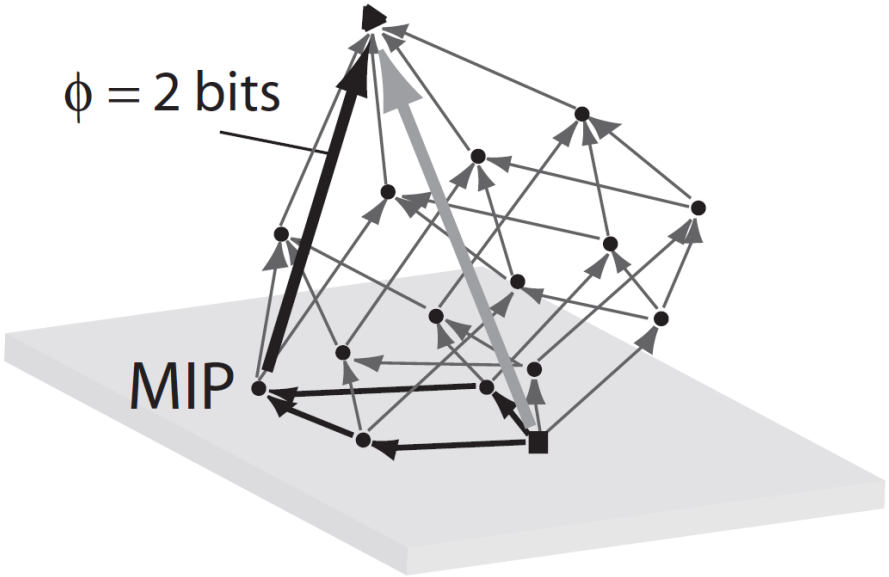
# Qualia as a shape in Q space

Very roughly:

- Imagine N dimensional space, where N is the number of possible states of the system
- Suppose system is in particular state (e.g. 1=ON, 2=OFF, 3=OFF, 4=OFF)
- Start with point (1/N,1/N…) e.g. complete uncertainty over N states.
- Add one connection at a time, recalculating repertoire until you have added all connections.
- Do for all orderings of connections.
- Connect all these the dots in the N dimensional space
- 'Height' of shape is Φ, 'shape' specifies the qualia.

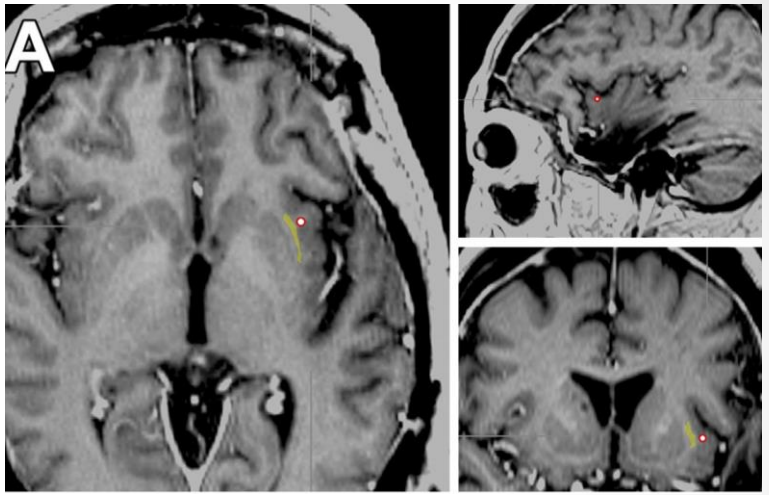
- See Tononi, 2008 for details!



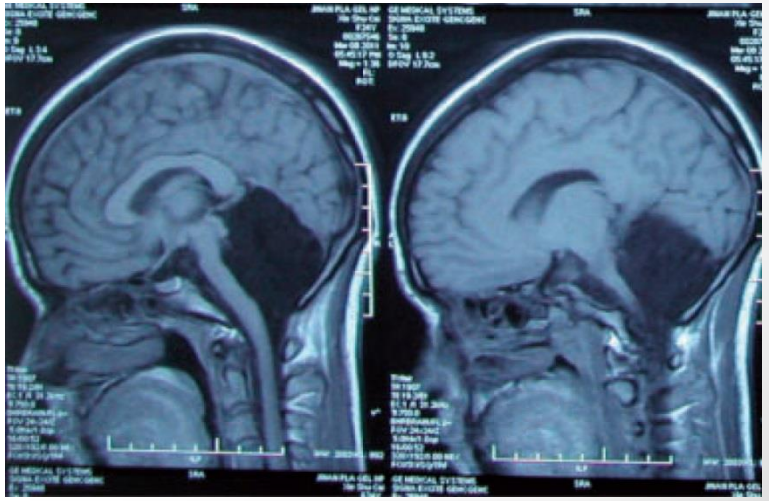



# Some soft neurobiological support

- Simulations suggest neural architecture of corticothalamic system, and especially claustrum likely to have high $\Phi$.
- Lesions to these areas serious impair consciousness (e.g. Koubeissi et al, 2014)
- Meanwhile, cerebellum has dense connections, but layered lattice like structure. Simulations show this likely to have minimal $\Phi$.
- A woman born in China with no cerebellum appears to have relatively unimpaired consciousness.
- See Koch's recent talk for many more examples: https://www.youtube.com/watch?v=LGd8p-GSLgY#t=1437




Caudate epilepsy lesion - Koubeissi et al, 2014



No cerebellum - Yu et al, 2014

# References


Balduzzi, D., & Tononi, G. (2008). Integrated information in discrete dynamical systems: motivation and theoretical framework. *PLoS Computational Biology*,*4*(6).

Balduzzi, D., & Tononi, G. (2009). Qualia: the geometry of integrated information. *PLoS computational biology*, *5*(8).

Block, N. (1980). Troubles with functionalism. *Readings in philosophy of psychology*, *1*, 268-305.

Chalmers, D (1995). Facing Up to the Problem of Consciousness. *Journal of Consciousness Studies* 2(3):200-19.

Edelman, G. M., & Tononi, G. (2013). *Consciousness: How matter becomes imagination*. Penguin UK.

Koubeissi, M. Z., Bartolomei, F., Beltagy, A., & Picard, F. (2014). Electrical stimulation of a small brain area reversibly disrupts consciousness. *Epilepsy & Behavior*, *37*, 32-35.

Levine, J. (1983). Materialism and qualia: The explanatory gap. *Pacific philosophical quarterly*, *64*(4), 354-361.

Oizumi, M., Albantakis, L., & Tononi, G. (2014). From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0. *PLoS Computational Biology*, *10*(5).

Nagel, T. (1974). What is it like to be a bat?. *The philosophical review*, 435-450.

Place, U. T. (1956). Is consciousness a brain process?. *British Journal of Psychology*, *47*(1), 44-50.

Putnam, H., (1960/1975). Minds and Machines. *Mind, Language, and Reality*, Cambridge: Cambridge University Press (p 362–385).

Shannon, C. E. (1948). Bell System Tech. J. 27 (1948) 379; CE Shannon. *Bell System Tech. J*, *27*, 623.

Smart, J. J. (1959). Sensations and brain processes. *The Philosophical Review*, 141-156.

Tononi, G. (2008). Consciousness as integrated information: a provisional manifesto. *The Biological Bulletin*, *215*(3), 216-242.

Yu, F., Jiang, Q. J., Sun, X. Y., & Zhang, R. W. (2014). A new case of complete primary cerebellar agenesis: clinical and imaging findings in a living patient. *Brain*.