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 $\Phi$
   3. $\Phi$ 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:
  o React to stimuli
  o Focus attention
  o Categorise/discriminate
  o Control behaviour

• ‘Hard’ problem –
  o Why are any of these processes accompanied by experience?
  o Why is there "something it is like" (Nagel, 1974) for us to discriminate the colour of a tomato?
  o 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

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

Extra baggage

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

Sensory inputs: e.g. standing on a nail

Mental state: e.g. pain

Behavioural outputs: e.g. jumping up and down
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)
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:
  o **The computational level** – What is the purpose of the system? What inputs should be mapped to what outputs?
  o **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?
  o **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.
Functionalism -> Computationalism

• Resurgence of functionalist ideas about mental states along with “cognitive revolution”

• Computationalism (e.g. Edelman, 2008; Tononi, 2008):
  o The brain/mind is an information processing system
  o Thinking is a form of information processing
  o 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


- 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

Uncertainty

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

$$H(\mathcal{X}) = -\left(\frac{1}{2} \log \frac{1}{2} + \frac{1}{2} \log \frac{1}{2}\right) \text{ bits}$$
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

Component ON (dark) if 2 or more of its causes were ON at t-1

State of system at time t, 1 is ON (dark) 2,3,4 OFF (light)

Effective information, due to constraints of system

The only possible state of system at t-1 given its structure

Possible states of system at t-1 ignoring its structure

$ei(X(\text{mech}, x_1)) = 4 \text{ bits}$

actual: $p(X_0(\text{mech}, x_1))$

potential: $p(X_0(\text{maxH}))$
Example system - Irreducible information

\[ p^{(a\text{M}_0(\text{mech,}\mu_1))} \]

\[ p^{(b\text{M}_0(\text{mech,}\mu_1))} \]

\[ \text{ei}^{(a\text{M}(\text{mech,}\mu_1))} = 1.1 \text{ bits} \]

\[ \text{ei}^{(b\text{M}(\text{mech,}\mu_1))} = 1 \text{ bit} \]
Example system

Comparing the ‘effective information’ of whole system (top), to product of MIP subsystems (bottom)

Product of distributions from previous page

\[ \phi(X(\text{mech}, x_1)) = H[p(X_0(\text{mech}, x_1)) \mid \prod_{k=1,2} p^{(kM_0(\text{mech}, \mu_1))}] = 2 \text{ bits} \]
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 $(1/N, 1/N, \ldots)$, 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 dots in the N dimensional space.
- ‘Height’ of shape is $\Phi$, ‘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 seriously 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
References


