The Science of Consciousness and Scientific Theories about the Brain

D A V I D G A M E Z

Department of Computer Science, Middlesex University, UK.

ToCAI Seminar, Middlesex University, London

4th March 2016

Talk Overview

- Scientific theories of consciousness.
- Scientific theories of the brain.

Scientific Theories of Consciousness

Current State of Consciousness Research

- Lot of experimental work on the neural correlates of consciousness.
- Some mathematical theories.
- But discussions about consciousness regularly return to philosophical problems (zombies, colour inversion, causation, hard problem etc.).
- These are legitimate concerns that could undermine scientific results.

Objectives

- An approach to consciousness that is
  - Mathematical.
  - General.
  - Free from philosophical problems.

My Approach

- Definition of consciousness that distinguishes it from the invisible physical world.
- Framework of definitions and assumptions that neutralizes the philosophical problems with the measurement of consciousness.
- Reject computational and informational theories of consciousness.
**My Approach**

- Obtain data by measuring:
  - Conscious states
  - Physical states.
- Use machine learning to discover mathematical relationships between descriptions of consciousness and descriptions of the physical world.
- These can answer questions about the consciousness of bats, brain-damaged patients and robots.

**First Part of Talk**

- Definition of consciousness.
- Science of consciousness:
  - Measurement of consciousness.
  - Measurement of the physical world.
  - Theories of consciousness.
  - Computational discovery of theories of consciousness.
- Applications

**Bubbles of Experience**

- In our everyday encounters with the world we are immersed in a bubble of space centred on our bodies.
- I call this a *bubble of experience*.
- Bubbles of experience contain colours, smells, sounds, etc.
Bubbles of Experience and the Physical World

- Over the last 400 years there has been a gradual change in our interpretation of bubbles of experience.
- Our theories about bubbles of experience have evolved in response to our theories about the physical world.

Naive Realism

- We are naive realists in our day-to-day lives.
- Colours, smells, sounds etc. are properties of objects.
- Objects continue to have these properties when they are not being perceived.

Naive Realism

Atomism

- Atomism was developed by the Ancient Greeks and revived in the 17th Century.
- Physical world consists of atoms and the void.
- Leads to a distinction between primary and secondary qualities:

Primary and Secondary Qualities

- Primary qualities:
  - Size, shape and motion.
  - Properties of atoms.
- Secondary qualities:
  - Colours, smells, sounds, etc.
  - Appear when atoms interact with the senses.
Bubbles of Experience and the Brain

Today the brain is interposed between our bubbles of experience and the physical world:
- The physical world interacts with the senses.
- Spiking patterns appear in the brain.
- These spiking patterns are somehow linked to bubbles of experience containing colours, smells, etc.
- No longer any reason to believe that our bubbles of experience resemble the physical world.

Modern Physics

In modern physics the physical world consists of wave-particles, superstrings, etc.
- We cannot imagine what these are like.
- The physical world has become a black box that is a source of signals (Russell).
- It no longer make sense to ask what the physical world is like.

What is Consciousness?

- "Consciousness" is another name for bubbles of experience, which have been reinterpreted in relation to the invisible physical world described by science.
- Fits in with Wilkes' claim that our modern use of the word 'consciousness' first appeared in the 17th Century.
Emergence of the Concept of Consciousness

Measurement of Consciousness
- Consciousness is measured through first-person reports.
- This raises a number of philosophical problems.
- These can be handled with assumptions.
- The science of consciousness is considered to be true given these assumptions.

CC sets
- A CC set is a set of spatiotemporal structures in the physical world that is correlated with a conscious state.
- This set is present when a conscious state is present and absent when the conscious state is absent.

CC sets are Functionally Connected to Consciousness
- A correlation between A and B is the same as a functional connection between A and B – they are different ways of stating that A and B deviate from statistical independence.
- There is a functional connection between a conscious state and its corresponding CC set.
Assumptions

1. The consciousness associated with a human brain is functionally connected to its reports.

2. All conscious states associated with a human brain are available for report and all aspects of these states can potentially be reported.

3. The conscious state associated with a CC set nomologically supervenes on the CC set. In our current universe physically identical CC sets are associated with indistinguishable conscious states.

4. CC sets cause a person’s first person reports about consciousness.

Assumptions for the Measurement of Consciousness

- Consciousness cannot be described in natural language:
  - Context-bound
  - Ambiguous
  - Not applicable to infants, bats, robots, etc.
  - Not mathematically tractable.

- Need a precise formal way of describing consciousness that is applicable to any system.

- This will be referred to as a c-description.

- Possible methods include:
  - XML/LMNL
  - High dimensional qualia
  - Category theory
Measurement of the Physical World

- The measured object interacts with a calibrated object.
- Observe result and extract a number.

Description of the Physical World

- The number that is extracted through a measurement procedure is attributed to an object in the physical world.
- 3 metres was the **height of an elephant**.
- Objects are tightly defined in physics and chemistry.
- They are not tightly defined in biology.

P-description

- We want a science of consciousness that can make predictions about the consciousness of arbitrary systems (bats, robots, rocks etc.)
- A science of consciousness based on biological neurons will not be able to say anything about the consciousness of systems based on synthetic neurons.
- Need a precise formal description of the spatiotemporal structures that form CC sets.
- Will be referred to as a **p-description**.

Theories of Consciousness
Theories of Consciousness

- A c-theory is a mathematical description of the relationships between c-descriptions and p-descriptions.
- It can generate c-descriptions from p-descriptions.
- It can generate p-descriptions from c-descriptions.

Example: Tononi’s Information Integration Theory of Consciousness IIT

- IIT (Tononi 2008) is the closest thing to a c-theory that we have so far.
- A mathematical algorithm links a description of the physical world to a description of consciousness.
- A conscious state (a quale) is c-described using a high dimensional mathematical structure.

Discovering C-theories

Traditional Scientific Discovery

- Traditionally, people have identified regularities in the physical world (Newton, Einstein, etc.).
- We generally assume that physical regularities are simple enough to be found by humans.

C-theories are Potentially Complex

- We have little or no idea about which spatiotemporal structures form CC sets.
- The mathematical relationships between c-descriptions and p-descriptions could be simple.
- Or they could be thousands of pages of differential equations.
**Methodology for Complex C-theories**

- If we assume that there are simple relationships between c-descriptions and p-descriptions, we could waste a lot of time and effort looking for something that does not exist.
- Better to devise a methodology that can:
  - Identify simple relationships between c-descriptions and p-descriptions.
  - Identify complex relationships between c-descriptions and p-descriptions.
  - Prove that no relationships exist in the current data.

**Computational Discovery of C-theories**

- Use computers to identify relationships between c-descriptions and p-descriptions.
  - Measure consciousness and the brain.
  - Use machine learning to identify simple and complex relationships (or prove that no relationships exist).
- This could draw on previous work in the computational discovery of scientific knowledge.
- Could be prototyped on a simulated neural network.

**Applications**

**Deductions about Consciousness**

- The science of consciousness could develop c-theories that can reliably map between p-descriptions and c-descriptions.
- Use reliable c-theories to make deductions about the consciousness of:
  - Patients in vegetative and minimally conscious states.
  - Infants.
  - Bats, octopi, snails, etc.
  - Computers and robots.

**C-theory Deduces Machine’s Conscious State**
Modification and Enhancement of Consciousness

- A similar procedure can be used to modify and enhance our consciousness or to build an artificial consciousness:
  1. Generate c-description of a desired state of consciousness.
  2. Use c-theory to convert c-description into p-description of a CC set.
  3. Realize the CC set in a machine or your brain.

Applications

- Deduce how to build a chip that extends or repairs a person's consciousness.
- Deduce how to upload a copy of my consciousness into a computer.
- Deduce how to produce desired states of consciousness in the brain.

Upload the Consciousness into Computer

Transcendence

Conscious Machines

Ex_Machina

Chappie

Construction of a Conscious Machine

- Description of physical state that is linked to desired state of consciousness
- Description of desired state of consciousness
- P-description → C-theory → C-description
- 3D printing
- Physical computer or robot

Modifications of Consciousness
Limitations

- Our current techniques for measuring the brain have very low spatial and/or temporal resolution.
- We lack formats for c-description and p-description.
- We are a very long way from a decent theory of consciousness.
- The science of consciousness and the technology of brain modification have a long way to go before we will be able to create states of consciousness on demand.

Theories of the Brain

- Lots of detailed information about the biology of neurons.
- Large-scale measurements of the brain (fMRI, EEG).
- Small-scale measurements of small numbers of neurons embedded in much larger networks.
- Links between genetics and large-scale behaviours (schizophrenia, Alzheimer’s, etc.)

Theories of the Brain

- The current science of the brain has weak predictive powers.
- Little idea about how the neurons work together to produce particular behaviours (sleep, talking, etc.)
- Can use known correlations to make some inferences based on brain measurements in fMRI scanner.
- No general theory that can make strong testable predictions.

Theories of the Brain

- What should a theory of the brain look like?
- What are neuroscience and experimental psychology hoping to achieve?
- Are we just going end up with a vast amount of unintelligible disconnected data?

Outline of Second Part of Talk

- Limitations of current theories about the brain.
- Generalize c-theories into b-theories.
- Computational discovery of b-theories.
- Conclusion
**Limitations of Brain Theories**

**Theories in Physics**
- Small sets of mathematical equations describe regularities in wave-particles, fields etc.
  - Newton’s equations (motion and gravitation).
  - Maxwell’s equations (electromagnetism).
  - Einstein’s field equations (gravitation and spacetime).

**Theories in Biology**
- Description and classification (Linnaeus).
- Genetics:
  - Very small number of bases (ACGT).
  - Different combinations of these bases lead to different organisms and some of their different behaviours.

**Reasons why the Brain is Tricky**
- Lack of access:
- Complexity and amount of data.
- Non-deterministic system – a given behavioural output can be produced by many different combinations of neural firing patterns.

**Biological Theories of the Brain**
- Descriptive theories:
  - Detailed information about how neurons work.
  - Correlation between brain area activation and behaviour.
  - Classification of neuron types.
- Genetic theories:
  - Link between genes and behaviour.
  - Link between genes and diseases (Parkinson’s, schizophrenia, etc.)

**Limitations of Biological Theories**
- Brain has a large number of very similar units (neurons).
- Neuron behaviour is only partly genetically determined.
- Few of the structures that are linked to behaviour are stable and consistent enough to be described by biological theories.
Cognitive Theories of the Brain

- Boxes representing memory, emotion etc.
- Mapped onto brain areas or functions.
- Often ‘tested’ by parameterizing the model to match a given set of experimental data.

Limitations of Cognitive Theories

- Boxes are subjective and badly defined.
- Weak predictive powers.
- Difficult to map functions down to brain areas that carry out multiple functions.

Mathematical Theories of the Brain

- Predictive coding (Friston 2005).
- Bayesian theories.
- Information integration (Tononi 2008).

Predictive Coding

- Activity in the nervous system is a process of matching internally generated predictions to external stimulation.
- Brains are hierarchically organized layers of predictive coders.
- Errors in prediction are passed up the hierarchy.
- Predictions are passed down the hierarchy.
- Lots of experimental work.
- Friston himself claimed that it could not be scientifically tested!

Bayesian Brain

- Brain represents knowledge of the world in terms of probabilities.
- Often claimed to be optimal according to Bayesian statistics.
- Lots of experimental research.
- The results are questionable (Bowers and Davis 2012).
Information Integration
- Mathematical algorithm used to identify:
  - The part of the brain that is conscious.
  - The amount of consciousness that is present.
  - The qualitative character of that consciousness.
- A good test for consciousness came out of work on information integration (Casali 2013), but this does not prove that the theory is correct.

Limitations of Mathematical Theories
- Information integration is not a theory of behaviour.
- Predictive coding and Bayesian theories are vague and difficult to test.
- Can a brain to come up with a theory about itself?

Another Way to Think About it
- Every brain is different.
- Brains change every day.
- A traditional scientific theory of the brain is going at best to be a theory of the average brain.

Scientific Theories of the Average Brain
- Describe what is common to all brains:
  - Details of individual neurons.
  - Activation of brain areas that process sensory data and other hardwired tasks.
- Most of the intermediate levels will be missing.
- Weak predictive powers.

My Approach to Consciousness
- Formal descriptions of conscious states.
- Formal descriptions of physical states.
- Mathematical theories of consciousness that map between these two formal descriptions.
**C-theory**

Generalizing this Approach

- Conscious reports are just one of the brain’s behaviours.
- Perhaps this approach could be generalized to all of the brain’s behaviours.
- Could be applied to very simple creatures that are probably not conscious.

**Generalizing the Approach**

- Formal descriptions of behaviour.
- Formal descriptions of brain states.
- Mathematical theories that map between these two formal descriptions.

**B-theory**

**Formal Description of Brain State**

- Would have to find the right level or combination of levels:
  - Neuron firing patterns.
  - Electromagnetic waves.
  - Chemistry.
  - Etc.
- Context would be less of an issue because we would not be seeking to generalize to non-biological systems.

**Formal Description of Behaviour**

- Differential or polynomial equations.
- Combination of sine, cosine, etc. to describe regular patterns in muscle contractions.
Mathematical Theory of Behaviour

- Would work in the same way as a mathematical theory of consciousness.
- Look for relationships between formal descriptions of behaviour and formal descriptions of consciousness.
- Should be able to generate testable predictions.

Discovering B-Theories

Computational Discovery of B-Theories

- B-theories might be simple
- Or they might be extremely complex (thousands of pages of differential equations).
- Better to start with a methodology that can find simple or complex b-theories (or prove that no regularities exist in the current data).

Type of Machine Learning

- The regularities discovered by machine learning should be transparent to the scientists.
- Weights in a deep neural network are not ideal.
- Polynomial equations or systems of differential equations are much better.

Computational Discovery of Physical Theories

Schmidt and Lipson (2009)
**Exploring this Approach**

- This approach could be tested on a simple biological system.
- For example, C. elegans or zebrafish larva.
- Could also be tested on a simulation of a simple biological system.

**C. Elegans**

- Nematode worm with 302 neurons.
- Connectome is known.
- Unclear how the 302 neurons interact to produce its behaviour.
- Unclear what a scientific understanding of the nervous system of C. elegans should be.

**OpenWorm Project**

- OpenWorm project could be a good place to start ([www.openworm.org](http://www.openworm.org)).
- Biologically realistic simulation of worm’s body and nervous system.

**Learning B-theory on C. Elegans**

- Three separate problems could be addressed with machine learning:
  - Compressed economical descriptions of the simulated neural network’s behaviour.
  - Compressed economical descriptions of the worm’s behaviour.
  - Identify mathematical relationships between the two descriptions.

**Human Brain Project**

- Could scale this up to more complex embodied neural network simulations.
- Eventually apply it to simulations of the brain.
- Human brain project is working towards the development of a working simulation of the brain.
- Some other promising work on scanning and brain uploading.

**Moving Beyond the Average Brain**

- Scientific theories of the average brain will always be weak.
- No one is going to sit down and study a single brain (if this were possible).
- Computational methods could be used to develop a b-theory that describes a single brain at a single point in time.
- This could have strong predictive powers.
Problems

- The approach to consciousness works (I think) because the spatiotemporal structures in the brain are linked to a single ‘thing’ (a bubble of experience) that is connected to first-person reports.
- States of consciousness are potentially stationary (few hundred milliseconds per state)
- It is not a behaviouralist theory – the mathematics is being applied to the relationship between conscious states and brain states.

Muscle Behaviour

- Muscle behaviour is temporally extended.
- Behaviours overlap.
- Hard to know when a single behaviour starts and stops.
- Might only be possible to develop a b-theory that applies to all of the organism’s possible behaviours.

... the era of simple mathematics effectively modelling parts of the world is drawing to a close. It is possible that new areas of investigation will lend themselves to simple models, but the evidence is that within existing areas of investigation, the domain of simple models has been extensively mined to the point where the rewards are slim.

Humphreys (2004)

Conclusions

Theories of Consciousness

- Consciousness defined as a bubble of experience and understood in relation to an invisible physical world.
- Framework of assumptions neutralizes the philosophical problems with measuring consciousness.
- Computational methods should be used to discover mathematical relationships between measurements of consciousness and measurements of the physical world.

C-theory
Theories of the Brain

- We do not have a clearly defined idea of what a theory of the brain should look like.
- I have suggested that a theory of the brain could be a mathematical relationship between:
  - Mathematical descriptions of neuron activity.
  - Mathematical descriptions of behaviour.

B-theory

- B-theories are likely to be extremely complicated (thousands of pages of differential equations).
- Could use machine learning to identify them automatically.
- This approach can be prototyped on simulations of simple and complex systems.

Acknowledgements

- Many thanks to Barry Cooper and the John Templeton Foundation for supporting this work (Project ID 15619: 'Mind, Mechanism and Mathematics: Turing Centenary Research Project').
- I would also like to thank Anil Seth, the Sackler Centre for Consciousness Science and the Department of Informatics at the University of Sussex for hosting me as a Research Fellow during this project.

References


More Information

- Slides: www.davidgamez.eu/talks
- Some papers related to this material: www.davidgamez.eu/publications
- Feel free to contact me if you have any questions: david@davidgamez.eu