Four Preconditions for Solving MC4 Machine Consciousness

DAVID GAMEZ
Middlesex University, London, UK

AAAI Spring Symposium: Towards Conscious AI Systems
25\textsuperscript{th} March 2019
Talk Overview

• Types of machine consciousness.
• Four problems:
  – Measurement of consciousness.
  – Description of the correlates of consciousness.
  – Description of consciousness.
  – Theories of consciousness.
• Conclusion.
TYPES OF MACHINE CONSCIOUSNESS
Types of Machine Consciousness

• **MC1.** Machines with the same external behaviour as conscious humans.

• **MC2.** Computer models of the correlates of consciousness.

• **MC3.** Computer models of consciousness.

• **MC4.** Machines that really have conscious experiences.
Types of Machine Consciousness

• **MC1.** Machines with the same external behaviour as conscious humans.

• **MC2.** Computer models of the correlates of consciousness.

• **MC3.** Computer models of consciousness.

• **MC4.** Machines that really have conscious experiences.
Conscious Human Behaviours

- Humans have characteristic behaviours when they are conscious.
- For example:
  - Alertness.
  - Response to novel situations.
  - Inward execution of sequences of problem-solving steps.
  - Learning.
  - Response to verbal commands.
  - Delayed response to stimuli.
MC1 Machine Consciousness

• A machine is MC1 conscious if it is producing similar external behaviour to a conscious human.
• Many artificially intelligent machines are already MC1 conscious to some extent.
• For example, humans can only play Atari video games, Go or Jeopardy! when they are conscious.
• MC1 machine consciousness is part of artificial general intelligence (AGI).
Types of Machine Consciousness

• **MC1.** Machines with the same external behaviour as conscious humans.

• **MC2.** Computer models of the correlates of consciousness.

• **MC3.** Computer models of consciousness.

• **MC4.** Machines that really have conscious experiences.
Models of the Correlates of Consciousness

- Researchers often model parts of the brain that are thought to be linked to consciousness (correlates of consciousness).
- MC2 machine consciousness is the construction of:
  - Models of the neural correlates of consciousness.
  - Models of the cognitive correlates of consciousness.
NeuroBot

- Neural implementation of global workspace.
- Controlled an avatar in the Unreal Tournament 2004 game environment.
- 20,000 neurons; 1.5 million connections.
- Implemented by Zafeirios Fountas.
Network Architecture

[Diagram showing the network architecture with nodes labeled as Sensory Module, Workspace Modules, Motor Module, Action Selection Module, Desired State, and Proprioception.]

Incoming sensory data flows through Sensory Modules, leading to Workspace Modules. These modules interact with Motor Modules, generating Motor output. The Desired State and Proprioception are also indicated in the diagram. The network architecture illustrates the flow of information and interactions between different modules within the system.
Unreal Tournament 2004

25/03/2019

David Gamez - Four Preconditions for Solving MC4 Machine Consciousness
Types of Machine Consciousness

- **MC1.** Machines with the same external behaviour as conscious humans.
- **MC2.** Computer models of the correlates of consciousness.
- **MC3.** Computer models of consciousness.
- **MC4.** Machines that really have conscious experiences.
MC3 Machine Consciousness

- Long tradition of describing the structure of consciousness from a first-person perspective.
- For example, Husserl and Merleau-Ponty.
- Can create computer models of conscious experiences in a machine.
- This is MC3 machine consciousness.
Imagination with CRONOS and SIMONS
Types of Machine Consciousness

- **MC1.** Machines with the same external behaviour as conscious humans.
- **MC2.** Computer models of the correlates of consciousness.
- **MC3.** Computer models of consciousness.
- **MC4.** Machines that really have conscious experiences.
Human Consciousness

- I define consciousness as a bubble of experience.
- A bubble of space roughly centred on our bodies containing smell, body sensations, colour etc.
Objective View
Human Consciousness

Representation of person's first-person experience of their body
MC4 Machine Consciousness

• A physical robot is MC4 conscious if it is associated with a bubble of experience.
• Its bubble of experience will contain something analogous to our colours, smells etc.
Conscious Machine (MC4)
Significance of Research on MC4 Machine Consciousness

• Ethical issues.
• Curiosity.
• We want to achieve immortality.
• Medical applications.
• Helps us to develop general scientific theories of human consciousness.
MC4 Conscious Machines
MC4 Consciousness
Transfer / Uploading

Transcendence

BLACK MIRROR

ALTERED CARBON
Solving MC4 Machine Consciousness

• MC4 machine consciousness would be solved if we had a *general* scientific theory of consciousness that could be applied to *any* physical system.

• To solve MC4 machine consciousness we need:
  – Clarity about the measurement of consciousness.
  – Generalizable ways of describing the correlates of consciousness.
  – Better ways of describing consciousness.
  – Mathematical theories of consciousness.
Solving MC4 Machine Consciousness

• MC4 machine consciousness would be solved if we had a *general* scientific theory of consciousness that could be applied to *any* physical system.

• To solve MC4 machine consciousness we need:
  – Clarity about the measurement of consciousness.
  – Generalizable ways of describing the correlates of consciousness.
  – Better ways of describing consciousness.
  – Mathematical theories of consciousness.
Measurement of Consciousness

• We cannot directly measure another person’s consciousness.
• Have to measure consciousness through external behaviour (first-person reports).
• This only works when we believe the system is capable of consciousness:
  – We assume that a system is capable of consciousness.
  – Then we use the system’s external behaviour to decide when it is actually conscious.
Inference from External Behaviour in Artificial Systems

• Some people believe that external behaviour can be used to infer the presence of consciousness in artificial systems.

• If this was the case, MC4 machine consciousness would be solved!

• Any system that exhibited particular external behaviours would be judged to be conscious.

• Inference from MC1 to MC4 machine consciousness.
Convincing Human-like Robots
Problems...

• Easy to write a computer program that mimics first-person reports about consciousness.

• A given set of external behaviours can be produced by a wide range of systems – giant lookup table, population of China communicating with radios and satellites, etc.

• Any sequence of physical states can be interpreted as a computer program that produces a given set of external behaviour.
Generalizable Theories of Consciousness

- Have to develop theories of consciousness by studying humans.
- Then *generalize* our theories of consciousness to other systems.
Generalizable Theories of Consciousness

• To make generalizable scientific theories of consciousness we need:
  – Generalizable ways of describing the correlates of consciousness.
  – Better ways of describing consciousness.
  – Mathematical theories of consciousness.
DESCRIPTION OF THE CORRELATES OF CONSCIOUSNESS
Description of the Correlates of Consciousness

• Promising data on the neural correlates of consciousness (particular brain areas, recurrent connections, HOT zones, etc.).
• This cannot be generalized to systems with different brains, such as birds, insects, cephalopods, etc.
• Cannot be generalized to artificial systems.
What is a Neuron?

• We do not have a precise definition that could tell us if an arbitrary piece of physical matter contains a neuron.
• Suppose we synthesize a neuron from basic biological components. At what point does it become a neuron?
• What about silicon neurons?
Generalizable Physical Correlates of Consciousness?

• We need a way of defining the correlates of consciousness that enables us to generalize from biological to non-biological systems.
• Could develop more precise definition of a neuron that applies to insects and synthetic neurons.
• Or focus on patterns in well-defined physical things – molecules, electromagnetic waves.
DESCRIPTION OF CONSCIOUSNESS
Limitations of Natural Language

• Contents of consciousness are often described in natural language, which has many problems:
  – Vague.
  – Compressed.
  – Context-dependent.
  – Human-centric assumptions about the nature of objects.
  – We only have words for human experiences.

• Cannot use natural language to describe the conscious states of artificial systems (or bats!).
Possible Solutions

• Need to find generalizable ways of describing states of consciousness.
• In earlier work I suggested how a markup language, such as XML, could be used to describe conscious states.
• Balduzzi and Tononi put forward a solution based on high-dimensional mathematical structures.
THEORIES OF CONSCIOUSNESS
Mathematical Theories of Consciousness

• Most plausible type of generalizable theory is a mathematical relationship between formal descriptions of the physical world and formal descriptions of consciousness.

• We can only develop this type of theory when we have figured out how to describe consciousness and the physical world.
Mathematical Theory of Consciousness

Description of physical state

P-description

C-theory

C-description

Description of conscious state

Measurement

Physical brain

C-report

Human

Four Preconditions for Solving MC4 Machine Consciousness
Deducing the MC4 Consciousness of a Machine

1. Scientific measurement
2. Description of physical state of artificial system
   P-description: $pd_8$
3. C-theory
   Reliable mathematical theory of consciousness
4. C-description: $cd_8$

Description of artificial system's conscious state
Building a MC4 Conscious Machine

1. Description of conscious state that we want to realize in the machine

2. C-theory
   - Reliable mathematical theory of consciousness

3. 3D printing, neuromorphic chips, etc.

4. Physical computer or robot
   - Machine contains CC set that is linked to specified conscious state

P-description: $pd_7$

C-description: $cd_7$
Mathematical Theories of Consciousness?

• Tononi’s information integration theory is a (unidirectional) example of a mathematical theory of consciousness.
• Could use machine learning to discover better mathematical theories of consciousness.
CONCLUSION
Conclusion

• MC4 machine consciousness: creation of machines that are associated with bubbles of experience.
• Many applications of MC4 machine consciousness research.
Conclusion

• MC4 machine consciousness will be solved when we have:
  – Clarity about the measurement of consciousness.
  – Generalizable ways of describing the correlates of consciousness.
  – Better ways of describing consciousness.
  – Mathematical theories of consciousness.
More Information

- Read for free, download and purchase at: https://www.openbookpublishers.com/product/545.
- Website with papers: www.davidgamez.eu.
Questions?

• Website: www.davidgamez.eu.
• Book: www.openbookpublishers.com/product/545
• Contact: david@davidgamez.eu.