Generalizable Scientific Theories of Machine Consciousness

David Gamez

Middlesex University, London, UK

Birbeck Research Seminar, 21st November 2018

Talk Overview

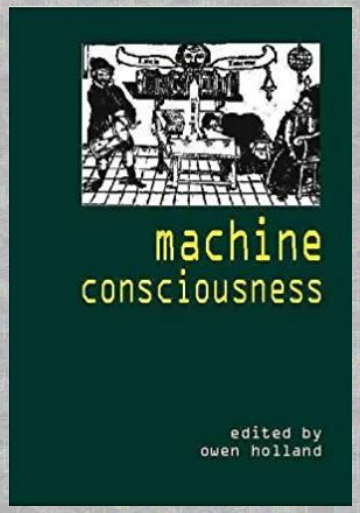
- Machine consciousness.
- MC4 conscious machines.
- Four problems:
  - Measurement of consciousness.
  - Description of the correlates of consciousness.
  - Description of consciousness.
  - Theories of consciousness.
- Conclusion.

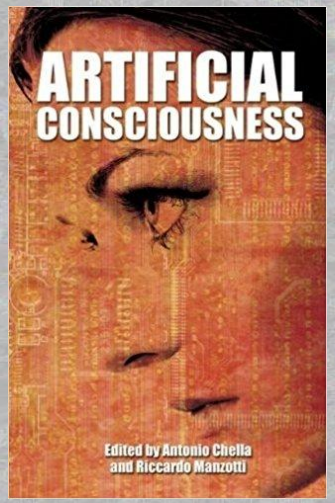
MACHINE CONSCIOUSNESS

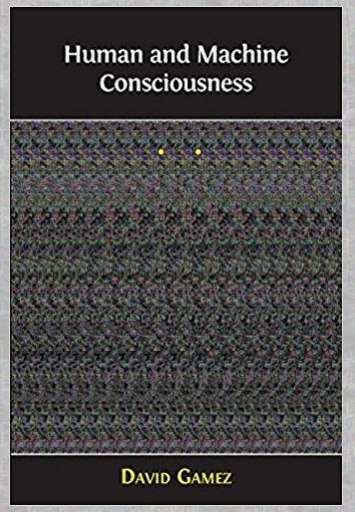
Machine Consciousness

- 1980s – revival of interest in consciousness.
- 1990s - serious scientific work on consciousness.
- Early 2000s – emergence of research on machine consciousness, sometimes called ‘artificial consciousness’.
- 2004 – Holland and Troiscanko’s awarded £500,000 to build a conscious robot.

Some Publications on Machine Consciousness







Gamez, D. (2008). Progress in Machine Consciousness. *Consciousness and Cognition* 17(3): 887-910.

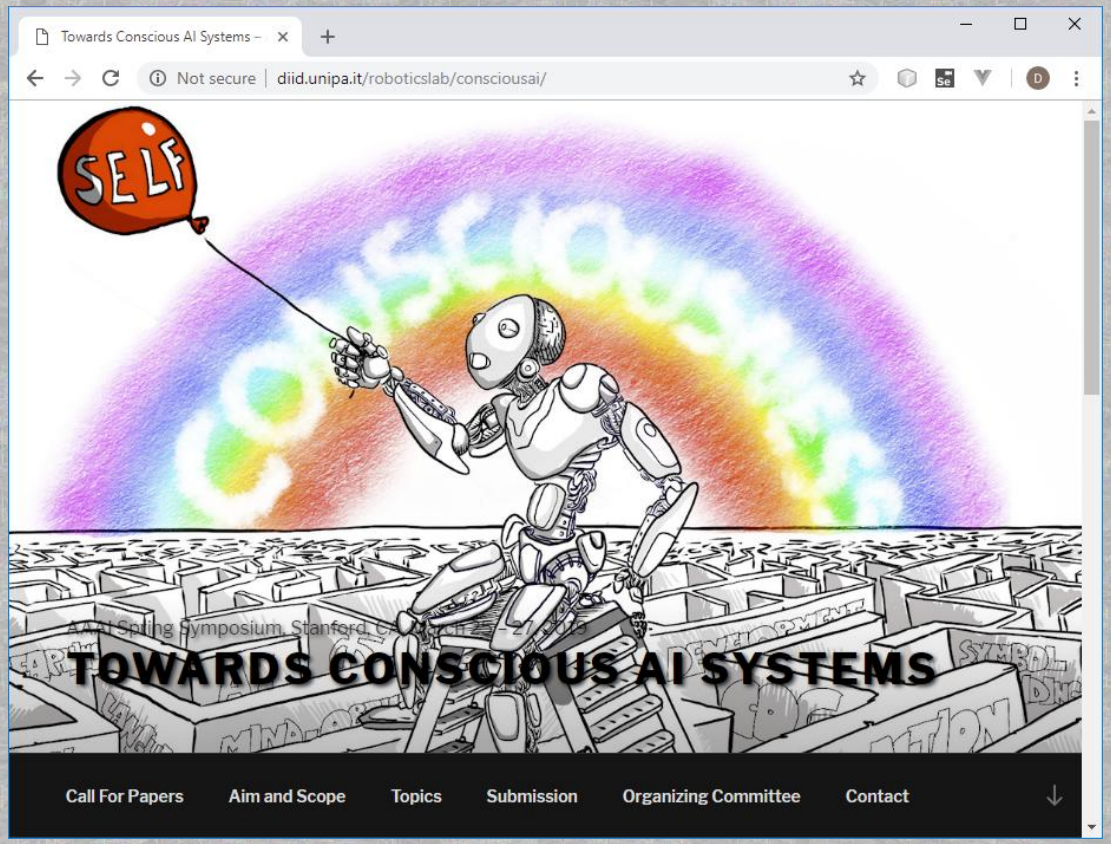
Reggia, J. A. (2013). The rise of machine consciousness: studying consciousness with computational models. *Neural Networks* 44: 112-31.

AAAI Symposium on ‘Towards Conscious AI Systems’

Website: http://diid.unipa.it/roboticslab/consciousai/

Palo Alto, USA.

27-27 March 2-19.



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).

IBM Watson


PIENSE
THINK
ΣΚΕΨΟΥ
DENKE
PENSER
$300,000
$1,000,000
$200,000
KEN
WATSON
BRAD

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).
- Helps us understand how consciousness might work.
- Might help us to build more intelligent machines.

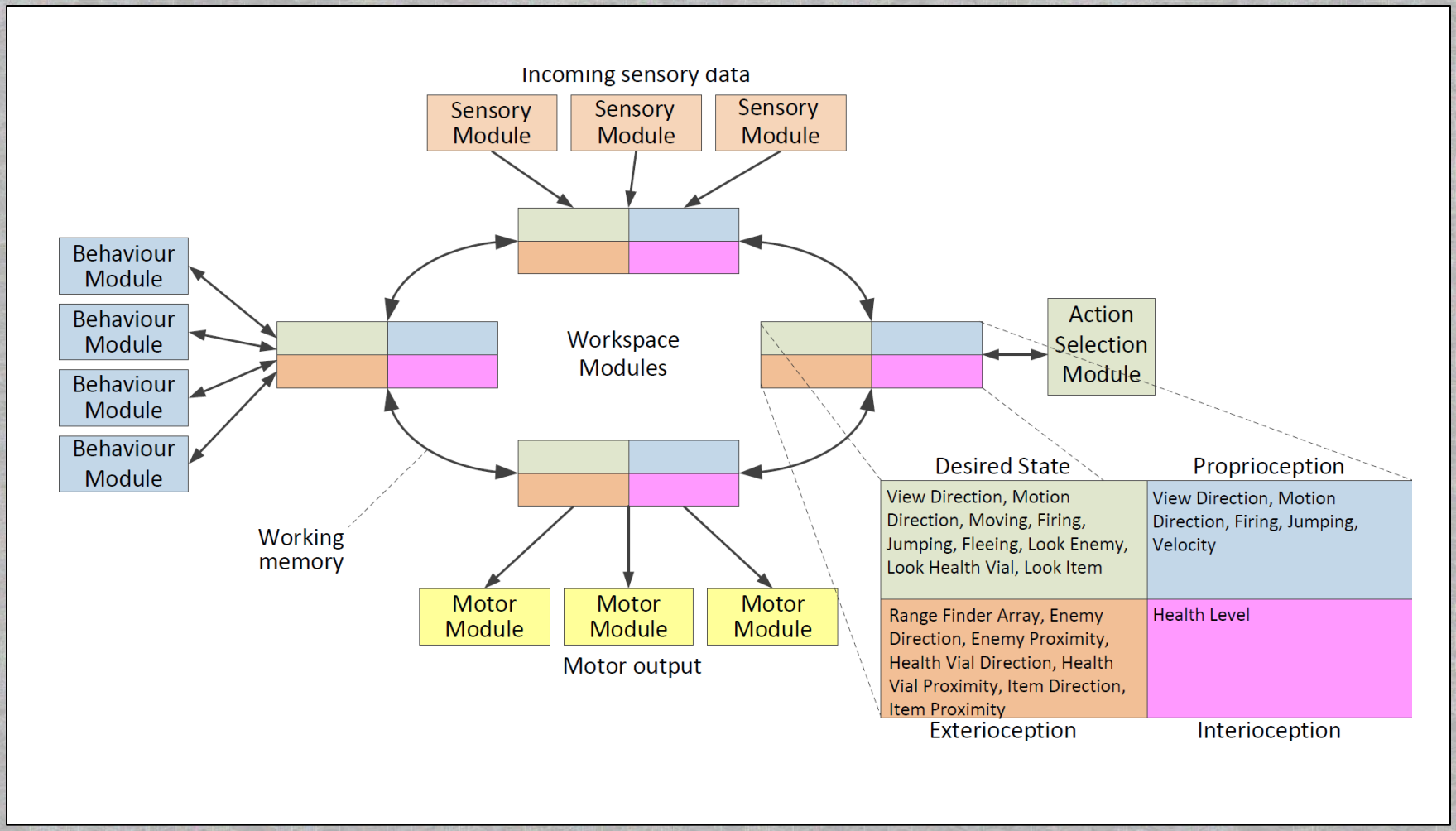
MC2 Machine Consciousness

- MC2 machine consciousness is the construction of:
  - Models of the neural correlates of consciousness.
  - Models of the cognitive correlates of consciousness.

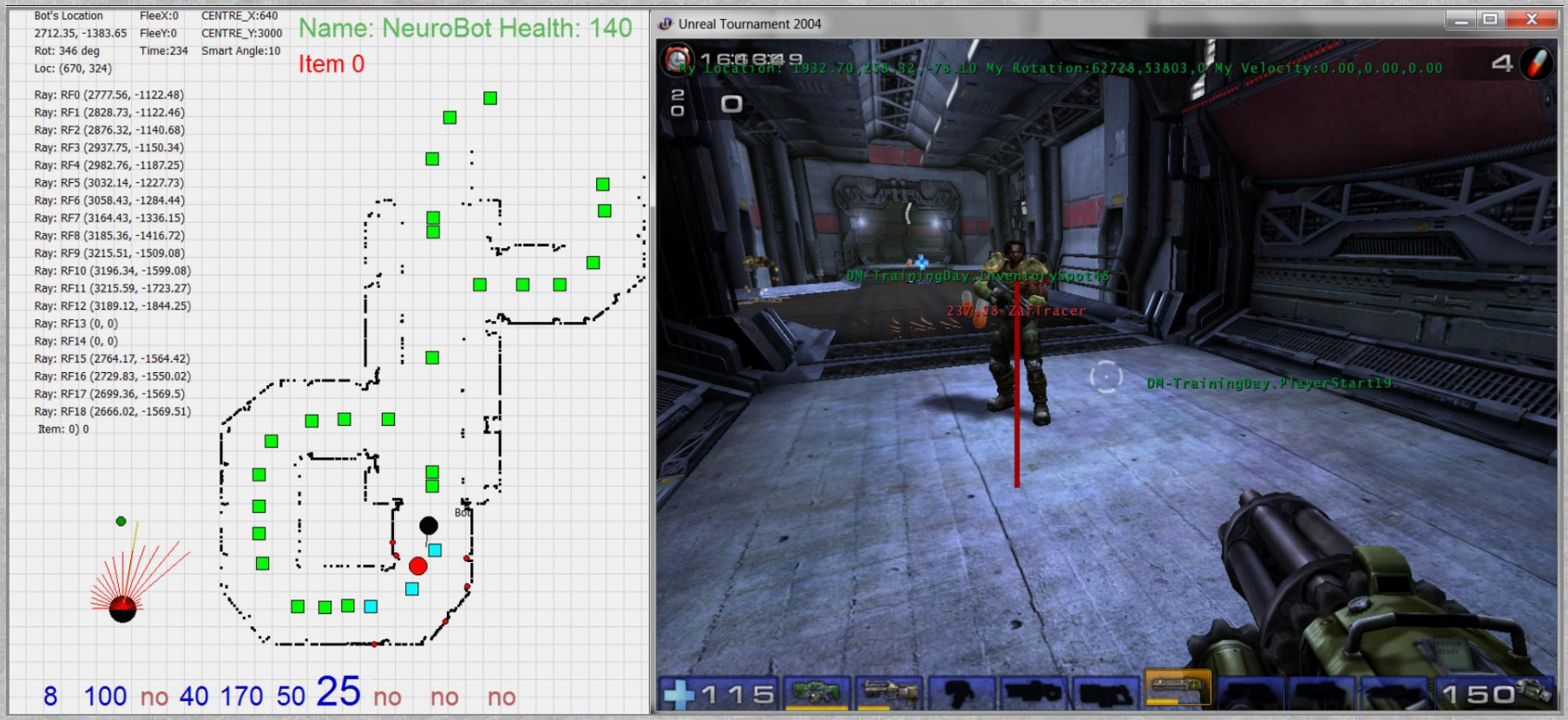
NeuroBot

- 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



Unreal Tournament 2004



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.

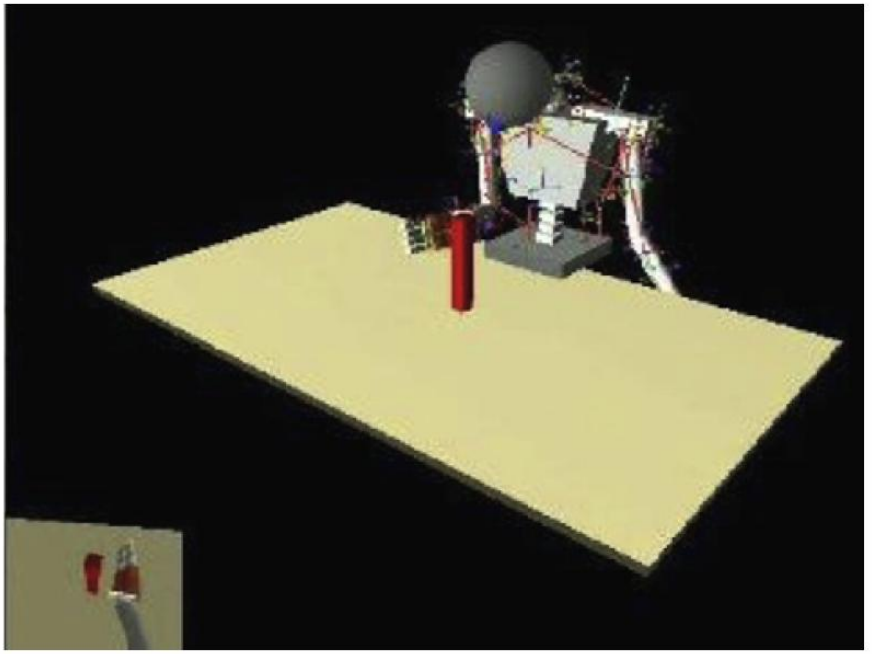
MC3 Machine Consciousness

- 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 SIMNOS

- Physical CRONOS robot controlled by a virtual model (SIMNOS).
- With SIMNOS the robot could 'imagine' different ways of solving a problem.
- When it found a solution, it executed it on the real physical robot.

Imagination with CRONOS and SIMONS



Types of Machine Consciousness

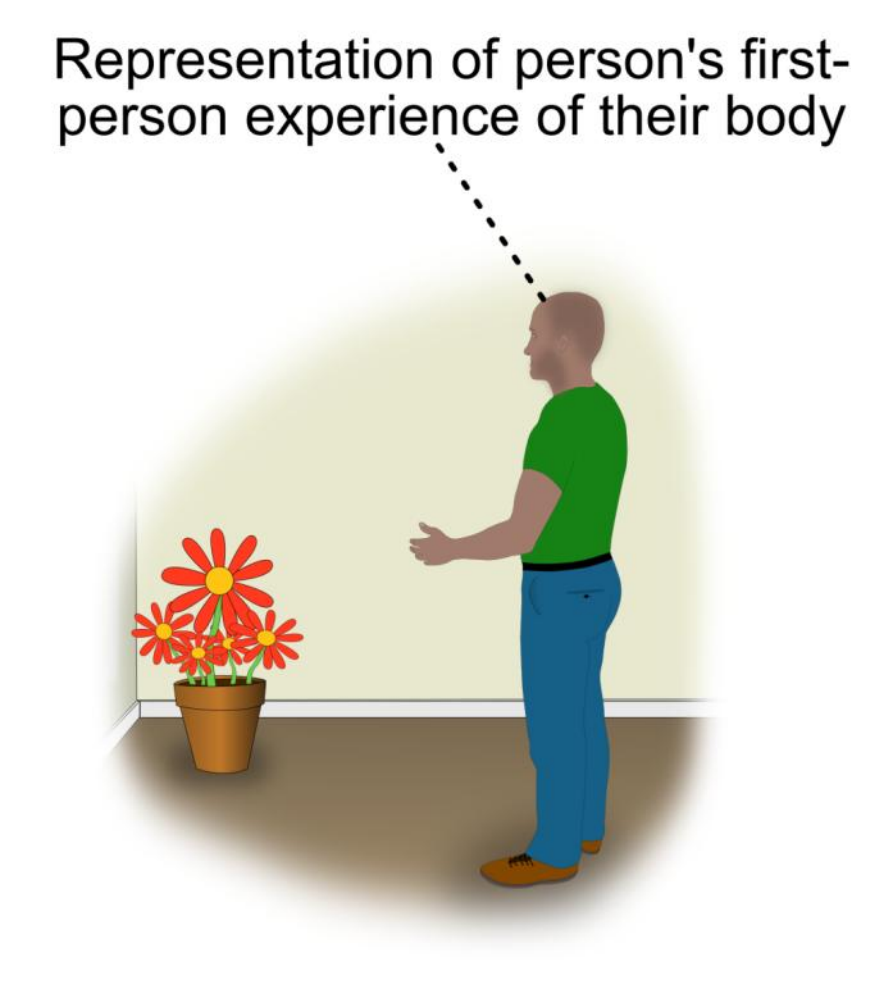
- **MC1**. Machines with the same external behaviour as conscious humans.
- **MC2**. Computer models of the correlates of consciousness.
- **MC3**. Computer models of consciousness.
- **MC4**. Machines that really have conscious experiences.

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

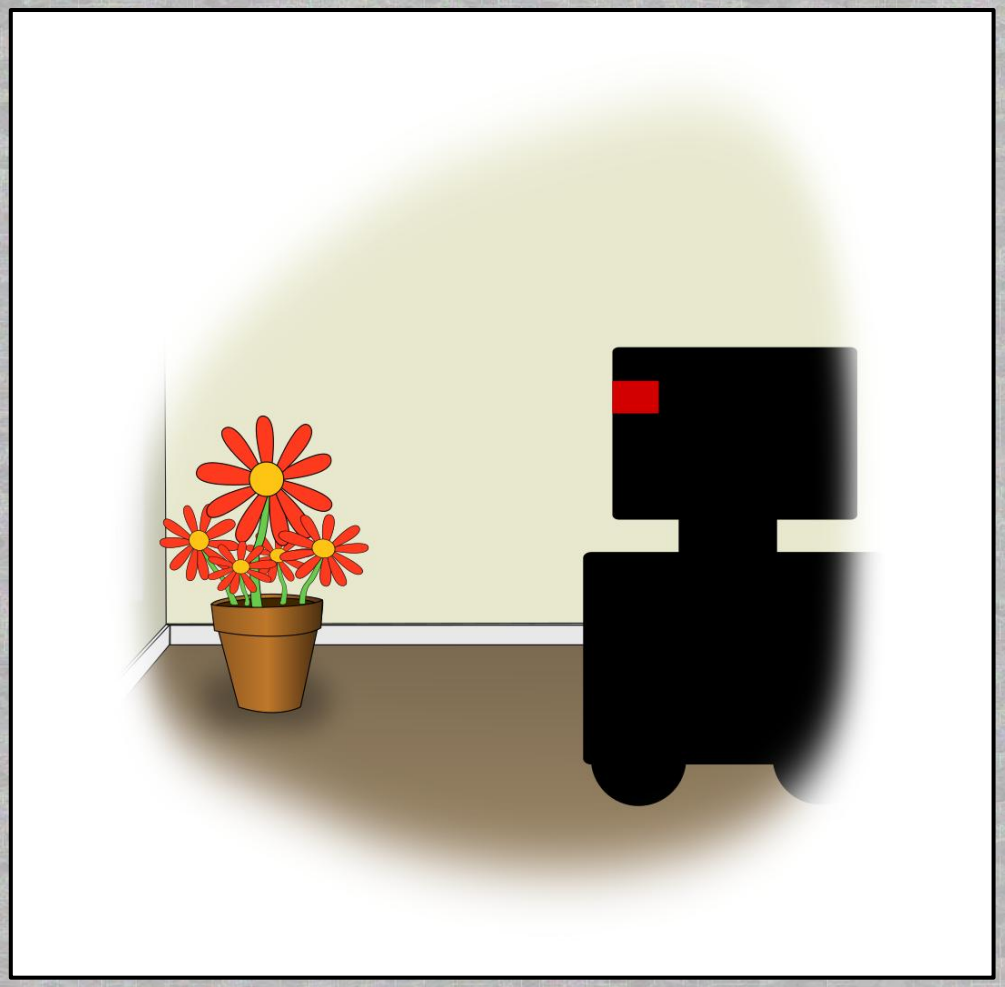
Human Consciousness



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)



Combinations of Types

- Different types of machine consciousness can be combined.
- A robot controlled by a model of the correlates of consciousness (MC2) could produce human-like behaviour (MC1).
- A robot controlled by a model of consciousness (MC3) could be associated with a bubble of experience (MC4).

MC4 Conscious Machines

MC4 Machine Consciousness

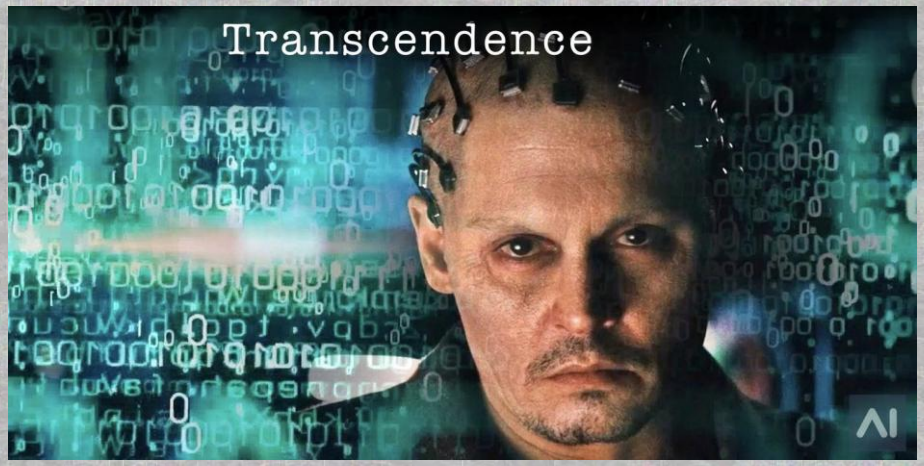
- When we are awake our brains are associated with bubbles of experience.
- A MC4 conscious machine is associated with a bubble of experience.
- A MC4 conscious machine has experiences that are analogous to our experiences of heat, colour, sound and pain.

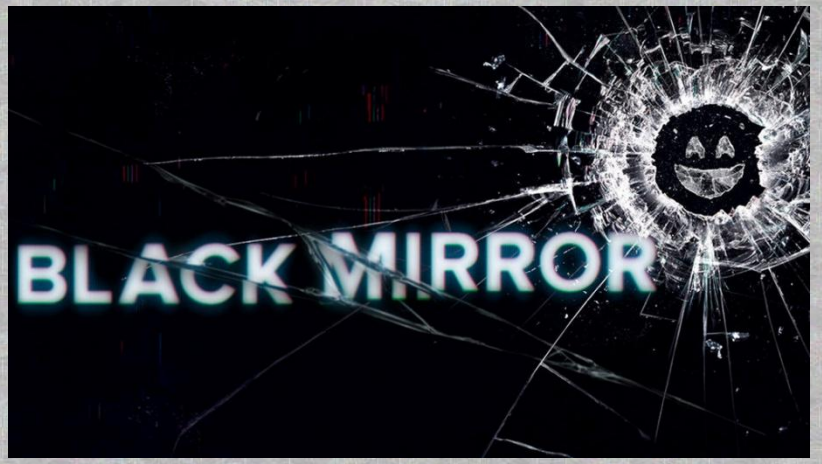
Significance of Research on MC4 Machine Consciousness

- 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





How to Solve MC4 Machine Consciousness

- MC4 machine consciousness could be solved if we had a *general* scientific theory of consciousness that could be applied to *any* physical system.
- We could measure the physical state of a machine and decide if it has the kind of stuff that is linked to consciousness or not.
- As simple as the question whether a machine contains carbon or not.

Current Scientific Theories of Consciousness

- Our current scientific theories of consciousness can't be generalized to machines until we have solved four problems:
  1. Measurement of consciousness.
  2. Description of consciousness.
  3. Description of the correlates of consciousness.
  4. Theories of consciousness.

Measurement of Consciousness

Measurement of Consciousness

- We cannot directly measure another person's consciousness.
- Have to measure consciousness through external behaviour (first-person reports).
- For example:
  - Delayed response to stimulus.
  - Flexible response to novel situations.
  - Inward execution of sequence of problem-solving steps.

Inference from External Behaviour

- 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 our science of consciousness by studying humans.
- Then we apply theories of consciousness to other systems.
- To make this work, we need to develop theories that are *generalizable*.

Generalizable Theories of Consciousness

- To make scientific theories of consciousness generalizable we need to solve problems with:
  - Description of the correlates of consciousness.
  - Description of consciousness.
  - 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 applied to:
  - Birds, insects, cephalopods, etc.
  - Artificial systems.

What is a Neuron?

- Neurons are defined inside a particular biological context – the brains of animals.
- Within this biological context there is considerable variation – size, morphology, etc.
- We do not have a precise definition that could tell us if an arbitrary piece of physical matter contains a neuron.

What is a Neuron?

- Suppose we genetically engineer a sequence of hybrids between neurons and liver cells.
- First cell is 100% neuron; middle cell is 50% neuron 50% liver; last cell is 100% liver.
- How do we classify intermediate cells in this sequence?
- Suppose we synthesize a neuron from basic biological components. At what point does it become a neuron?
- What about silicon neurons?

Generalizing with Computational and Functional Theories

- We need a way of defining the correlates of consciousness that enables us to generalize from biological to non-biological systems.
- Many people have attempted to solve this problem by looking for functional or computational correlates of consciousness.

Problems with Functionalism and Computationalism

- Functionalism leads to panpsychism.
- No workable theory of computational implementation. So we cannot prove that there are functional or computational correlates of consciousness.
- Current theories of computational implementation can find any given function in both the conscious and unconscious brain.

How can We Describe the Physical Correlates of Consciousness?

- Abandon functional and computational theories of consciousness.
- Focus on patterns in well-defined physical things – molecules, electromagnetic waves.
- Develop more precise definition of a neuron that applies to insects and synthetic neurons.

Description of Consciousness

Description of Consciousness

- Majority of work on consciousness is based on contrastive analysis of conscious and unconscious brains.
- Consciousness is described as 1 or 0!
- Much less work on detailed relationships between contents of consciousness and physical states.

Description of Consciousness

- Contents of consciousness are typically described:
  - In natural language.
  - Using a physical stimulus.
- Neither method can be generalized to artificial systems.

Limitations of Natural Language

- Conscious states are typically described in natural language. This 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!).

Limitations of Stimulus-based Descriptions

- We often describe conscious contents using the stimuli that produced them.
- For example, in work on brain reading with fMRI, decoded conscious experiences are presented as videos, which the subject compares with their own conscious experiences.
- But the consciousness that I have when I view a video is extremely unlikely to be the same as the consciousness of a machine that views the same video.
- So videos and other stimuli cannot be used to describe the consciousness of artificial systems.

New Ways of Describing Consciousness

- We have to study consciousness in humans and then generalize to artificial systems (measurement problem).
- Our method for describing consciousness in humans must be generalizable to radically different forms of consciousnesses.

Possible Solutions

- In earlier work I suggested how a markup language, such as XML, could be used to describe conscious states.
- Balduzzi and Tononi have put forward a solution based on high-dimensional mathematical structures.

THEORIES OF CONSCIOUSNESS

Theories of Consciousness

- There is a lack of general agreement about the final form that theories of consciousness should take.
- Generalizable theories should meet the following criteria:
  1. Generate testable predictions.
  2. Applicable to any physical system.
  3. Compact (Occam's razor).
  4. Based on objective properties of the physical world.

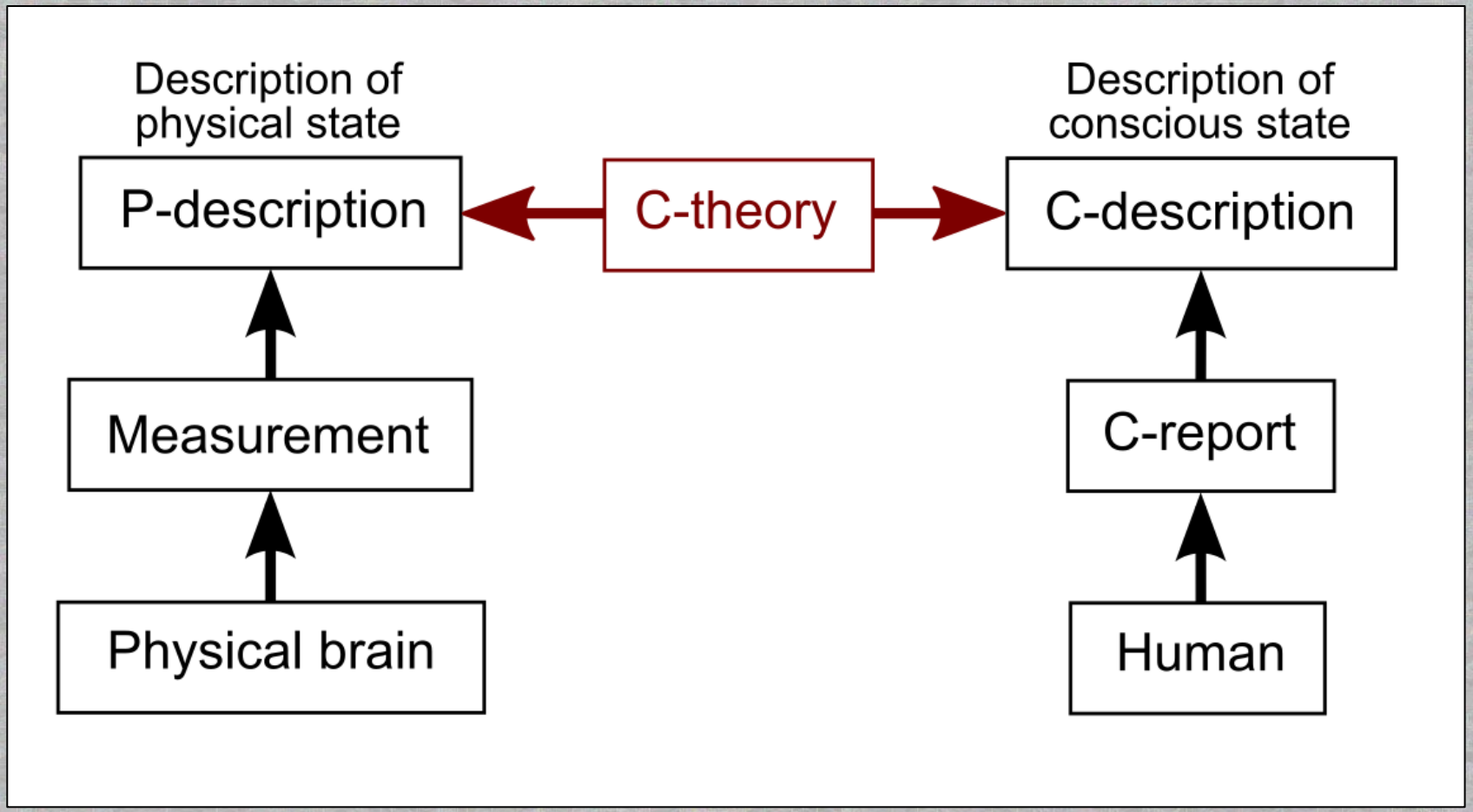
Unworkable Theories of Consciousness

- Functional, computational and informational theories of consciousness fail to meet criteria 4: They are not based on objective properties of the physical world.
- Theories based on neurons are not compact or generalizable and they have a weak ability to generate testable predictions.

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



Information Integration Theory of Consciousness

- Tononi's information integration theory of consciousness (IIT) is a good example of the final form that a theory of consciousness should take.
- Algorithm converts a description of the physical world into a high-dimensional mathematical structure that describes the level and contents of consciousness.
- Has some major limitations:
  - Based on information.
  - Can only be applied to systems with ~12 elements.

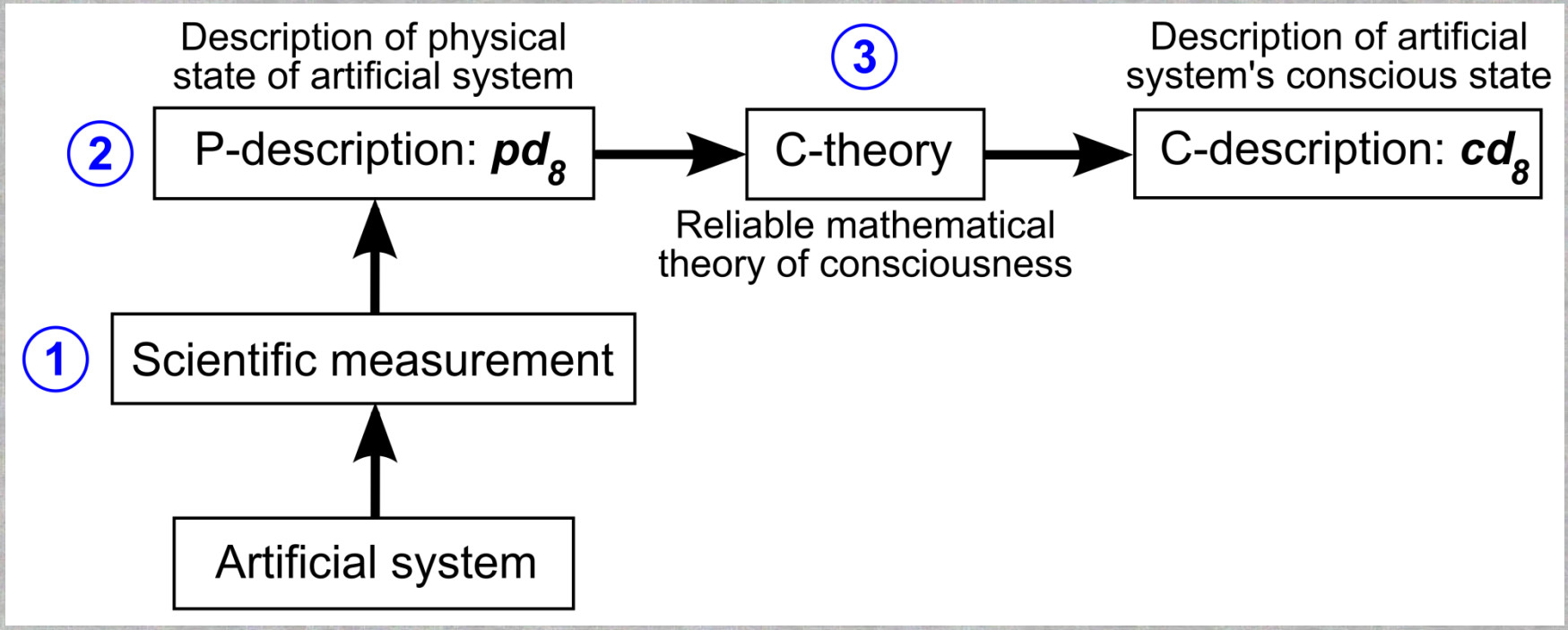
Mathematical Theories can Solve MC4 Consciousness

- When we have developed a reliable mathematical theory of consciousness we will be able to:
  - Generate believable predictions about MC4 consciousness in machines.
  - Build machines that are associated with specific conscious states.

Deducing the MC4 Consciousness of a Machine

- Measure physical state of machine.
- Use reliable theory of consciousness to convert description of physical state into description of conscious state.

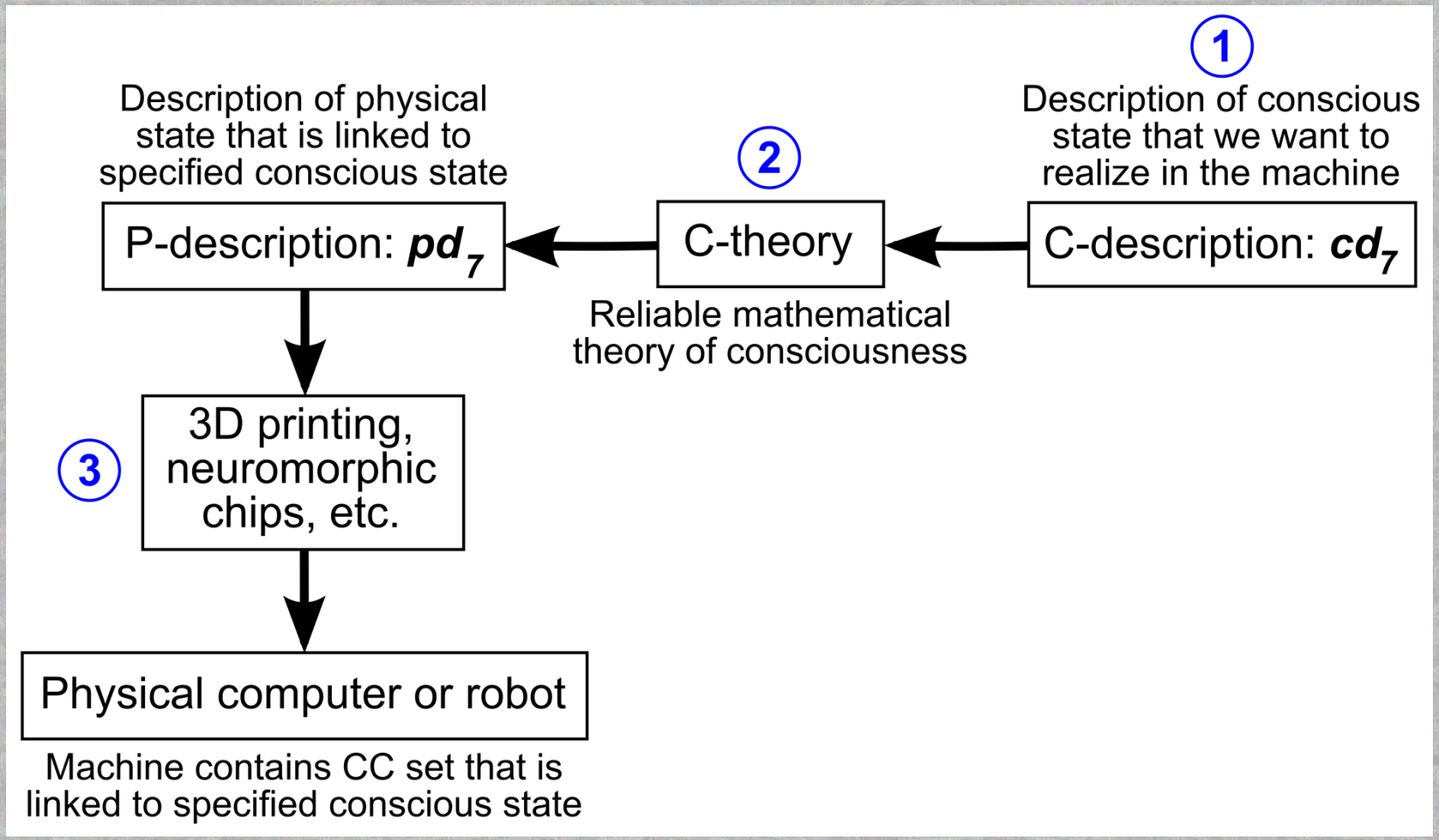
Deducing the MC4 Consciousness of a Machine



Building a MC4 Conscious Machine

- Generate description of the state of consciousness that we want in the machine.
- Use reliable theory of consciousness to convert description of consciousness into description of the corresponding physical state.
- Realise physical state in a machine.

Building a MC4 Conscious Machine



Discovering Mathematical Theories of Consciousness

- Mathematical theories of consciousness will be based on high resolution data from the brain.
- Potentially billions of pieces of data that cannot be comprehended by a human brain.
- Will have to use artificial intelligence to discover mathematical theories of consciousness.
- Could leverage previous work on computational scientific discovery.

CONCLUSION

Types of Machine Consciousness

- **MC1**. Machines with the same external behaviour as conscious systems.
- **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.

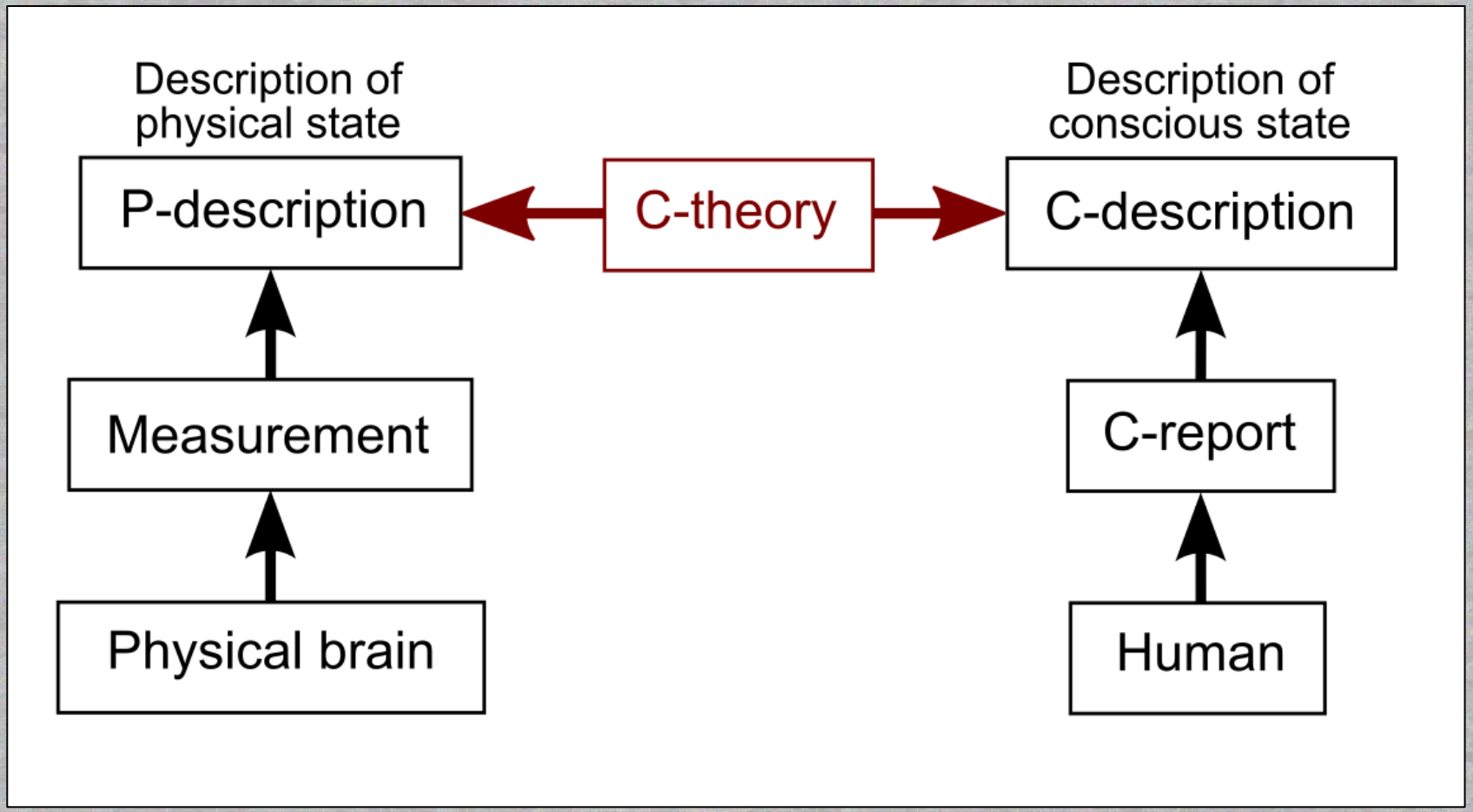
Significance of Research on MC4 Machine Consciousness

- Ethical issues.
- Curiosity.
- We want to achieve immortality.
- Medical applications.
- Helps us to develop general scientific theories of human consciousness.

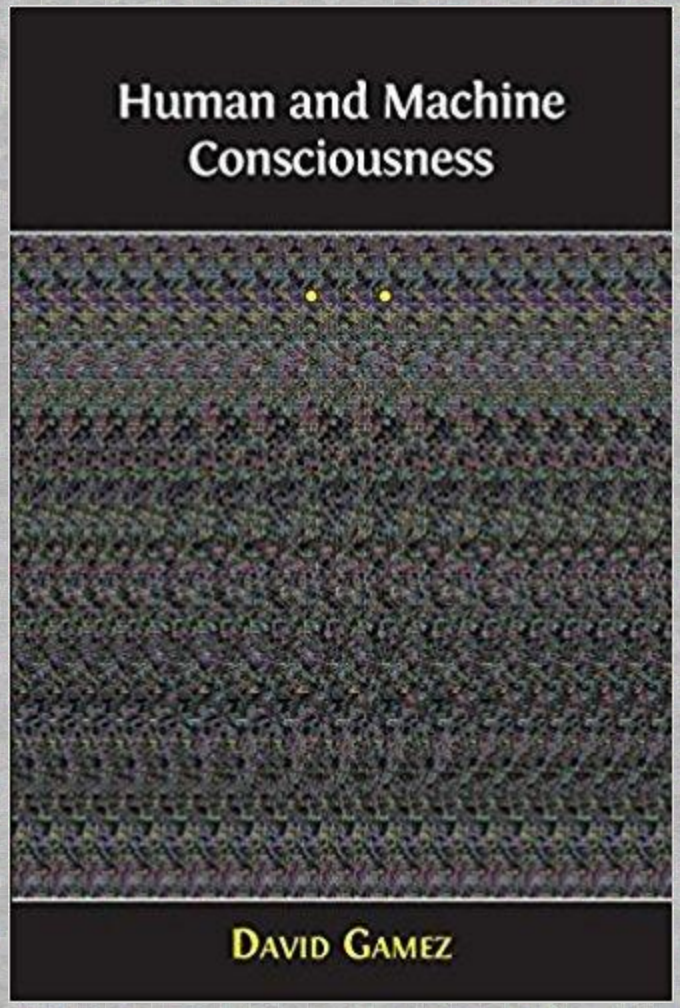
Current Scientific Theories of Consciousness

- Need to solve four problems before science of human consciousness can be generalized to MC4 machine consciousness:
  1. Measurement of consciousness.
  2. Description of the correlates of consciousness.
  3. Description of consciousness.
  4. Theory of consciousness.

Mathematical Theory of Consciousness

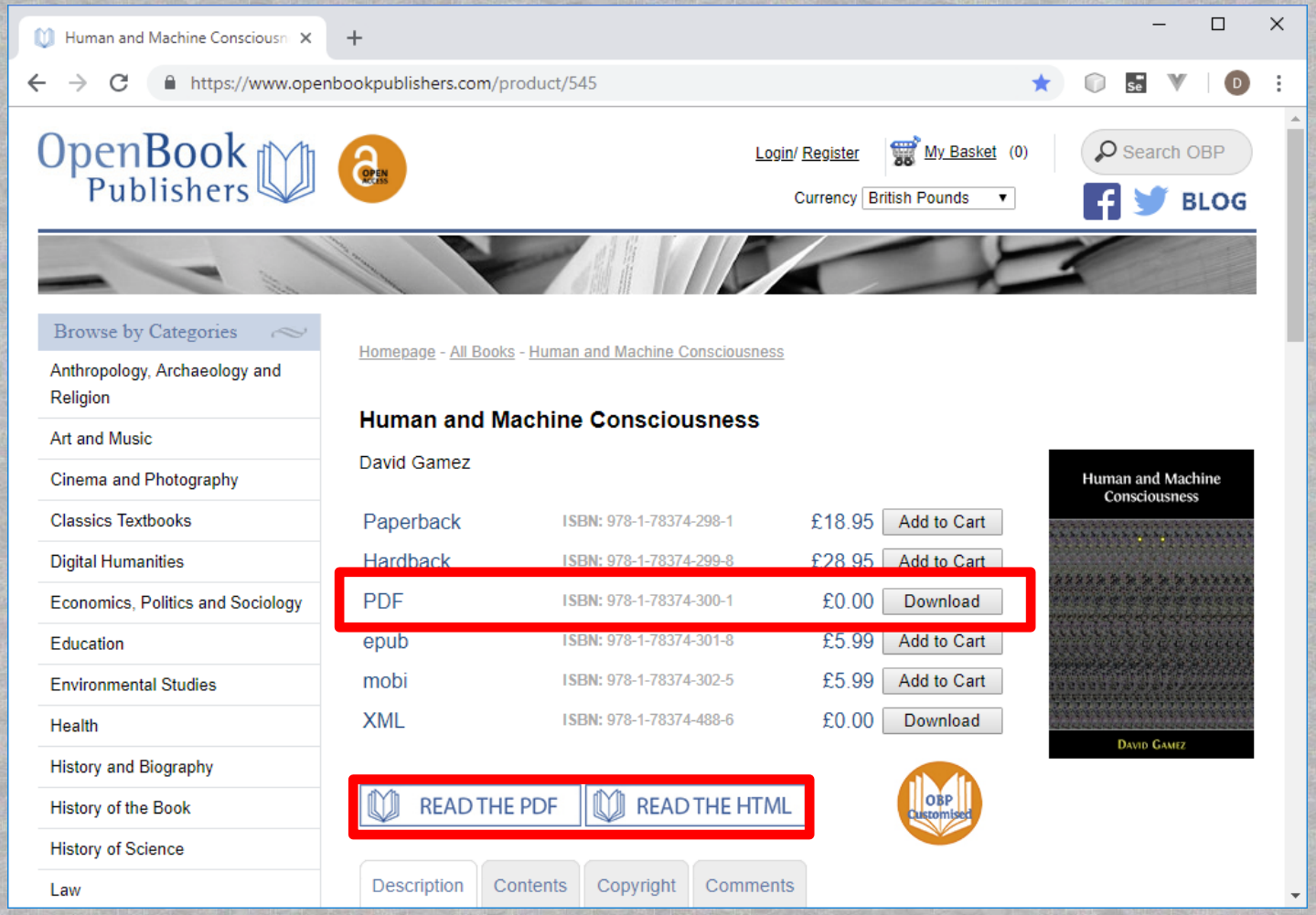


More Information



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

Human and Machine Consciousness



AAAI Paper

Four Preconditions for Solving MC4 Machine Consciousness

David Gamez[1]

[1] Department of Computer Science, Middlesex University, London, NW4 4BT, UK
d.gamez@mdx.ac.uk / www.davidgamez.eu

Abstract. A machine is MC4 conscious if it has phenomenal experiences that are comparable to human conscious experiences. From an ethical point of view it is important to know whether we have created MC4 consciousness in a machine. MC4 consciousness research can also contribute to the development of general theories of human consciousness. This paper discusses four problems that have to be solved before we will be able to address MC4 machine consciousness in a systematic way: We need more clarity about the measurent of consciousness, we need better ways of describing the physical world and consciousness, and we need to reach agreement about the final form that a theory of consciousness should take. When these problems have been addressed we will be able to develop scientific theories of consciousness that can make accurate believable predictions about MC4 consciousness in machines.

Keywords. consciousness, machine consciousness, artificial consciousness, science of consciousness, neural correlates

1 Introduction

It is often helpful to distinguish four types of machine consciousness [1]:

- **MC1**. *Machines with the same external behavior as conscious systems*. Humans behave in particular ways when they are conscious. For example, they are alert, they can respond to novel situations, they can inwardly execute sequences of prob-

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 from 2012-2015.

Questions?

- Website: www.davidgamez.eu.
- Contact: david@davidgamez.eu.