Engineering consciousness

One organiser's summary of the ESF PESC Exploratory Workshop: *Models of Consciousness*, Birmingham, September 1-3 2003

Ron Chrisley

Director, COGS

University of Sussex

Overview

- Principles of the engineering approach to consciousness
- Definitions of consciousness mentioned at the MoC meeting
- Tests for consciousness
- Other issues arising

Engineering consciousness (EC)

- Primary goal: Construction of machine consciousness
- But need a theory of consciousness to guide design (and thus construction)
- So should consider a theory's ability to guide design when constructing and evaluating it
- But design and construction can also inform theory

Constraints on theory

- Design
  - At least: Implementable
  - Better: Suggests designs
- Data
  - At least: Consistent with
  - Better: Explains
- Other
  - Simplicity
  - Unifiability with other theories

EC: Sources of data

- Phenomenology (e.g., entire visual field is coloured)
- Phenomenological reports (e.g., subjects report that entire visual field is coloured)
- Behaviour (other than phenomenological reports; e.g., grip of blindsight subjects)
- Architectural constraints (conceptual; e.g., fear that bridge might collapse requires ability to entertain counterfactuals)

EC: Operationalisation (1)

- If we just think of our goal as building a conscious machine, it can be difficult to see how to proceed, or to tell if we have succeeded
- Instead, we can "operationalise" the notion of consciousness:
- Why are we interested in making a conscious machine, anyway? What do we want it to do?

EC: Operationalisation (2)

- Interested in making a machine that has:
  - Autonomy
  - Adaptivity/advanced learning capacities
  - Emotion/affect
  - Responsibility (or to which we are responsible)
  - Intelligence
  - Authenticity (own world view and goals)
  - Ability to integrate information from different sources/modalities

EC: Operationalisation (3)

- Interested in making a machine that has:
  - Vivid/meaningful sensation/perception
  - Ability to act in the world
  - Ability to simulate/imagine/plan
  - Ability to represent its own states
  - Attentional capacities
  - A belief that it is conscious, gives phenomonological reports

EC: Implementability (1)

Non-implementable theories, that claim that consciousness is

- *Indefinable*
- *unknowable*
- *epiphenomenal* (in the sense of having *absolutely* no causal powers)
- *just a myth* invented by philosophers
- *not possessable* by (non-biological!) machines
- such that *machines indistinguishable from us could lack it*

are not useful for an engineering approach, and were (supposed to be!) off-limits during the MoC workshop.

EC: Implementability (2)

- Thus, on the design/engineering approach it will not count as an objection to a theory that one can imagine (in Chalmers' strong sense) something which satisfies that theory but is not conscious

EC: Theory evaluation

- What will count is a theory's stance on clear cases (positive and negative)
- It will count against a theory if it:
  - Implies that something isn't conscious which, pre-theoretically, is clearly conscious (e.g., humans)
  - Implies that something is conscious which, pre-theoretically, is clearly not conscious (e.g., a stone)
- Of course, what we take to be the "clear" cases can change

EC: Summary

- Primary goal: construction of artificial consciousness
  - Design informed by (and informs) theory
- Secondary goal: explanation of natural consciousness
  - So theory also constrained by data (especially clear cases), simplicity, unifiability, etc.
- Operationalisation
  - Thus, implementability is mandated (no zombies!)
- Complements other approaches

Consciousness is/involves (1)

- (Pre-reflective) Self (Taylor, Salichs)
- Transparency (Haikkonen, Sanz)
- Learning (of dynamics) (Salichs)
- Planning (Salichs)
- Heterophenomenology (Chrisley, Sloman)
- Split of attentional signal in a way that provides infallible self-identification (Taylor)
- Centralised action selection (Redgrave)
- Control of attention (Anceau, Edmondson)
- Timing management (Anceau)

Consciousness is/involves (2)

- Computational correlates (Cleeremans):
  - Meta-representation
  - Representational quality
    - Strength
    - Distinctiveness
    - Stability
- Accessibility (Cleeremans)
- Meaning generation (Sanz, Haikonen)
- Control (Sanz, Holland, Sloman, Chrisley)
- Affect/Motivation (Sloman, Manzotti)
- Information processing architecture (Sloman)
- Uniqueness/non-duplicability (Manzotti)

Consciousness is/involves (3)

- Imagination/simulation
  - Also pondering future, daydreams?
  - Chrisley, Hesslow, Doran, Ziemke, Holland, Revonsuo, Haikonen, Shanahan, Sanz
  - Not at workshop: Baars, Franklin, Stein, Dawkins?, et al.

Tests for consciousness? (1)

Lacombe (ESF): How could you tell if a machine were conscious?

- Posner attentional benefit task (Taylor)
- Priming, subception, deep discrimination (Booth)
- Seriality (Anceau)
- Operationalised tests (Chrisley)

Tests for consciousness? (2)

- Generating meanings (Sanz):
  - A system is aware if it is generating meanings from perceptions;
  - its degree of awareness is proportional to the proportion of the action space which it can model/anticipate
  - A system is conscious if “I am aware” is generated from the perceptual flow (thus, higher order theory?)
- There can be no single test for consciousness (Sloman)
- Other minds (Chrisley)

Tests for consciousness? (3)

Not at workshop:

- Pragmatics/Ethics (Brooks, Dennett)
- Does the machine do something that, in a human, requires consciousness? (adapted from Minsky?)
- Same as above, but not just same behaviour, but also same functional process

Other issues/questions

- Is consciousness a (single) scientific kind?
- Must conscious systems be Living/autopoietic/biological?
- Method: theory before, during, or after design of conscious artefacts?
- Is there an active/non-active distinction for consciousness?
- Prosthetic consciousness?