Architectural Requirements for Consciousness

Ron Chrisley
School of Engineering and Informatics
University of Sussex

Aaron Sloman
School of Computer Science
University of Birmingham

EuCognition2016, December 8-9th, Vienna

UNIVERSITY OF SUSSEX

CENTRE FOR COGNITIVE SCIENCE

the Sackler Centre for Consciousness Science

Introduction: A change of focus

- Original focus: "Qualia without dualism"
- Two parts:
  - Part 1: Philosophical argument:
    - The Hard Problem of consciousness is because of qualia
    - Qualia can be explained (not eliminated) ***if*** we can identify the features of our cognitive architecture that make use disposed to believe that there are aspects of our engagement with the world that are private, immediate, intrinsic and ineffable
    - Thus, whether there really are qualia or not may be an ***empirical*** matter

Introduction: A change of focus

- Part 2: Cognitive architecture design:
  - Identify some specific features of a cognitive architecture that would satisfy the conditions of the philosophical account above
    - E.g., the disposition to believe that one is in a state that one can have knowledge of directly, not by virtue of any other relation
  - Identify the general features of a cognitive architecture required for the specific features above

From our extended abstract

- *"These include those for being a system that can be said to have beliefs and propensities to believe.*
- *Further, having the propensities to believe 1-4 requires the possibility of having beliefs about:*
  - *oneself,*
  - *one's knowledge,*
  - *possibility/impossibility,*
  - *and other minds.*
- *At a minimum, such constraints require a cognitive architecture with:*
  - *Reactive,*
  - *Deliberative, and*
  - *Meta-management components (Sloman and Chrisley 2003)*
- *With at least two layers of meta-cognition:*
  - *(i) detection and use of various states of internal VM components; and*
  - *(ii) holding beliefs/theories about those components."*

Cognitive Architecture vs. *Cognitivist* Architecture?

- But given the (mostly healthy and reasonable) anti-representational, anti-symbolic, embodied, enactivist, etc. inclinations of many in our community, this might have been "burying the lead"
- Instead, change the focus to how one can have a grounded, dynamic, embodied, enactive(ish) cognitive architecture that supports the notions of belief, inference, meta-belief, etc.

Caveats

- This is likely not optimal/feasible/original?
- It is only intended to act as a proof-of-concept
- Slides only written over the past day or so, so no diagrams, no proper mathematical typesetting, etc.
- In particular: Have not checked with Aaron!

Robot and environment

- Consider a robot that can move its single camera to fixate on points in a 2D field
  - R(x,y) = s.
- The field is populated by simple coloured polygons, at most one (but perhaps none) at each fixation point (x,y).
- Suppose the field is static during trials, although it may change from trial to trial.

Feature map

- Suppose the robot has learned a discrete partition F of categories of S (e.g., a self-organising feature map):
  - M(s) = f_i in F.
- For example, f_1 might be active in those situations in which there is a green circle, f_2 might be active in those situations in which there is a red triangle, etc.

The task

- After a tone is heard, a varying cue (for example, a green circle) appears in some designated area outside of the field (the upper left corner say).
- The robot's task (for which it will be rewarded) is to perform some designated action (e.g. say "yes") if and only if there is something in the current array that matches the cue, that is:
  - "Yes" iff Exists (x,y): M(R(x,y)) = M(cue).

Strategy 1: Exhaustive search of action space

- There are several strategies the robot could use to perform this task.
- 1: It could perform a serial exhaustive search of the action space R(x,y), stopping to say "yes" if at any point M(R(x,y)) = M(cue).
- Time-consuming

Strategy 2: Exhaustive search of *virtual* action space

- Prior to hearing the tone, it can learn a forward model (E_w) from points of fixation (x,y) to expected sensory input (s) at the fixated location:
  - E_w(x,y) = s member of S
- It could then perform a serial exhaustive search of the Expectation space E_w(x,y), stopping to say "yes" if at any point M(E_w(x,y)) = M(cue)
- But suppose the task must be performed very quickly, more quickly than can be done even vi virtual exhaustive search

Strategy 3: Reflection

- First, for any given cue (nodes in the feature map), note that we can define the set P_cue to be all those parameter sets w that yield a forward model that contains at least one expectation to see that cue:
  - P_cue = All w: Exists (x,y): M(E_w(x,y)) = cue

Strategy 3: Reflection

- With a network distinct from the one realising E, the robot can learn an approximation of P_cue:
  - F_j(w, cue) = 1 iff w is in P_cue
- That is, F is a network that:
  - takes the parameters of E as input
  - outputs a 1 only if those parameters realise a forward model E_w for which there is at least one action (x,y) for which E_w expects to receive the given cue as input after performing R(x,y).

Strategy 3: Reflection

- Then the third way a robot might solve the task is to just input the current E parameter configuration w and the cue into F, and say "yes" iff F_j(w, cue) = 1.
- Why is this interesting?
- Because it is the beginning of reflection, of meta-belief.
- And that opens to door to inference, and sensitivity to logical relations.
- To see how, consider one more addition to this architecture

Conjunctive cues

- As with the individual nodes in the feature map, we can define the set P_c1,c2 to be all those parameter sets w that yield a forward model that contains at least one expectation to see c1 and one expectation to see c2:
  - P_c1,c2 = All w: Exists (x1,y1)(x2,y2)
    - M(E_w(x1,y1)) = c1 and
    - M(E_w(x2,y2)) = c2

Conjunctive Reflection

- With a network G distinct from E and F, the robot can learn an approximation of P_c1, c2:
  - G_k(w, c1, c2) = 1 iff w is in P_c1,c2
- That is, G is a network that:
  - takes the parameters of E as input
  - outputs a 1 only if those parameters realise a forward model E_w for which:
    - there is at least one action (x1,y1) for which E_w expects to receive an input that M maps to c1 after performing R(x1,y1); and
    - there is at least one action (x2,y2) for which E_w expects to receive an input that M maps to c2 after performing R(x2,y2)

Logic sets in

- Note that it is a logical truth that
  - If w is in P_c1,c2, then w is in P_c1
- So it is also true that
  - If G_k(w, c1, c2) = 1 then it should be the case that F_j(w, c1) = 1
- The robot could observe and learn this regularity

The awareness of inconsistency

- But because F and G are only approximations, there might actually be cases (values of w) where they disagree (where F(w)=0 but G(w)=1)
- To resolve this, should the robot modify F or G
  - Depends on the situation
- But the important point is that the robot has the essentials of a notion of logical justification and logical consistency of its own beliefs

Going meta

- But does this only work for basic-level beliefs (expectations to receive inputs), or might it apply to itself (beliefs about beliefs)?
- Proposal: give the same story as before, but the relevant expectations are not about motor actions and expected sensory input…
- But rather expectations about how one's sensorimotor forward model(s) will change if one alters them in this or that way
- Result: sensorimotor-grounded deliberation, meta-management, meta-belief, etc.

Thank you