A conversation with Professor Aaron Sloman, July 3, 2016

Participants

- Professor Aaron Sloman – Honorary Professor of Artificial Intelligence and Cognitive Science, University of Birmingham
- Luke Muehlhauser – Research Analyst, Open Philanthropy Project

Note: These notes were compiled by the Open Philanthropy Project and give an overview of the major points made by Professor Sloman.

Summary

The Open Philanthropy Project spoke with Professor Sloman of the University of Birmingham as part of its investigation into which types of beings should be of moral concern, and thus a potential target for the Open Philanthropy Project's grantmaking. This conversation focused on one particular factor plausibly relevant to whether a being should be of moral concern or not — namely, whether that being is phenomenally conscious, and what the character of its conscious experience is. Conversation focused on identifying potentially useful inputs for a "model combination / cluster thinking" method to produce rough predictions about which entities are conscious, and relevant work being done in related fields (e.g. philosophy, neuroscience, and artificial intelligence).

Professor Sloman asked to be referred to as "Aaron" in the remainder of this document.

Luke's approach and Aaron's approach

Luke explained that, given the Open Philanthropy Project's current uncertainty about the likelihood of any particular theory of consciousness being correct, he uses a "model combination" / "cluster thinking" approach to the question of which beings are likely to be conscious. This approach takes several types of evidence (e.g., predictions made by theories of consciousness, polls of experts, "table of features that might indicate consciousness" approaches) as inputs, weights them in some way, and outputs rough probabilities for the phenomenal consciousness of various beings (e.g. various species). Some possible inputs for a model combination approach are discussed below.

In contrast, Aaron explained that he tries to understand the variety of forms of consciousness produced by biological evolution on the basis of the information processing challenges of organisms of varying degrees of complexity and the kinds of awareness different biological designs provide. Some of that is awareness of current sensory input, some awareness of the immediate environment, some awareness of an extended environment, some awareness of past states and processes in the environment, and some of it is awareness of possible, likely, or impossible future situations. Those forms of awareness are externally directed. For many biological purposes internally directed awareness is also useful, and at different stages and lineages of evolution different sorts of self-awareness evolved.

Some of them are not fixed for a species, but depend on individual learning and needs. Some of the internally directed forms of awareness are concerned with current, past, and possible future sensory states. Others are concerned with current, past and possible future forms of reasoning, decision making, and learning. For example, if some mode of reasoning works in one situation and produces a bad or unexpected result in a later situation it is useful to be able remember the past reasoning process and try to identify previously unnoticed differences that might account for different results, e.g. a failed plan in one case and a successful plan in another. These are merely illustrative examples.

Aaron thinks that a survey of all the varieties of awareness, the different kinds of functions they may serve, the mechanisms required to support them, and the opportunities and difficulties involved in replicating them in computers is a long term research problem. For example, one important kind of human awareness allowed our ancestors to make discoveries leading up to the knowledge of geometry, topology and arithmetic in Euclid's *Elements* over 2000 years ago. But we are nowhere near giving current AI systems that kind of awareness. There are many other aspects of awareness that are "other directed" and evolve various kinds of evaluation (including self-evaluation).

Some of the information processing requirements, including the architectural layers, required in order to replicate such biological phenomena have been investigated in the Birmingham Cognition and Affect project since around 1991, building on work Aaron and colleagues did in earlier decades at Sussex University. But there are still huge gaps in our understanding of the forms of awareness that biological evolution has produced (built on by cultural evolution), and our abilities to model them. More information can be found at http://www.cs.bham.ac.uk/research/projects/cogaff/ and at http://www.cs.bham.ac.uk/research/projects/cogaff/misc/vm-functionalism.html.

Identifying and comparing phenomena that involve consciousness

Aaron thinks it might be useful to identify various phenomena that appear to involve consciousness and look for differences between them that might imply differences in their underlying mechanisms. For example, it might useful to identify which of humans' consciousness-involving abilities are or are not shared with other animals.

Examples of shared behaviors include:

- Seeing an object, moving towards it, and picking it up (an ability which humans have in common with, e.g., birds) appears to require at least those aspects of consciousness necessary for grasping spatial relationships and conceiving of ways of changing those relationships.
- Identifying a desired outcome and pursuing it via a series of movements that involve altering spatial relations between objects (e.g. moving twigs to build a nest) appears to involve choosing from a branching set of

potential action paths leading from the current state of the twigs to the desired state of the nest.

Work by ethologists on shared behaviors

Many ethologists have examined these types of abilities, though often without an explicit focus on the implications for consciousness. A common view among researchers in this area is that animals learn, by experience, probability distributions over possible outcomes (e.g. the probability that a certain action will result in a particular reward), and use these to choose future actions. Aaron thinks this particular model may be flawed and suggests that animals may also be able to use modal ideas (i.e. "possibility," "necessity," "impossibility") without having explicit probability distributions, as illustrated in this discussion document: http://www.cs.bham.ac.uk/research/projects/cogaff/misc/impossible.html.

Identifying differences between humans and other species

It might also be useful to identify ways in which human cognition differs at different stages of development, as well as how humans differ from other species. One example is human meta-cognitive abilities, i.e. the ability to think about one's own thought processes (though it is possible some other species can also do this). Meta-cognition allows individuals to notice, examine, and generalize particular cognitive methods that they use, which is useful for learning how to perform new tasks based on prior experiences involving similar but relevantly different tasks (e.g. manipulating objects to avoid a particular obstacle).

Aaron also suggests that mathematical discovery has often depended on thinkers' ability to apply a known cognitive process to a new situation, notice that it does not work as it did in previous situations, and modify the method appropriately.

Meta-cognition is also useful for modeling other minds, as made apparent in, e.g., false-belief tasks (such as the Sally-Anne test).

Poll of experts

Aaron thinks that a poll of experts' opinions on theories of consciousness could be useful as one input to a model combination approach (perhaps weighted relatively weakly). At present he is generally unimpressed by current theories of consciousness that are not based on analysis of (software) engineering design requirements for intelligent animals and machines. Merely observing and measuring what exists does not give as much insight as attempting to design and implement something similar. Unfortunately the education of researchers in most disciplines concerned with these problems does not usually include deep experience of computational modeling in a variety of paradigms. Currently available paradigms may have to be extended significantly, however.

Current approaches in related fields

Aaron's impression is that fields related to these issues (e.g. artificial intelligence, neuroscience, psychology) have by and large not attempted to explore in detail the mechanisms by which simple cognitive processes are implemented in humans and other animals. He is not aware of other researchers attempting to elaborate the details of human consciousness (by, e.g., proposing reasonably well-specified algorithms that might plausibly produce specific features of consciousness) in the same way that he is.

Aaron's impression is that most researchers who are trying to create models of how specific features of consciousness are produced have not incorporated certain data points that he considers relevant, e.g. the example of novel mathematical discoveries, discussed above. An example of this is the intuitive obviousness of certain topological facts, which seems to rely on deep cognitive processes that are not conscious and not well understood.

Global workspace theory

Aaron believes that some of the constraints implied by global workspace theory (GWT) do not fit with evidence about how human cognition works. For example:

- GWT does not seem to allow for concurrent cognition and meta-cognition (i.e. having an experience while also noticing and thinking about features of that experience), which humans appear to be able to do.
- GWT does not seem to allow for simultaneous inconsistent mental contents, which GWT proponents claim is empirically supported as a feature of consciousness (e.g. by the phenomenon of binocular rivalry). However, Aaron believes that this constraint is not consistent with some experimental situations that do appear to produce inconsistent mental contents in subjects (e.g. certain binocular rivalry setups that cause the subject to perceive different images from each eye simultaneously, superimposed).
- GWT does not seem to allow for simultaneous conscious awareness of the current world-state and of potential ways to alter that world-state.

Daniel Dennett

Some of Daniel Dennett's positions are consistent with Aaron’s approach, although Dennett does not explore some of the details of cognitive processes that are of interest to Aaron (e.g. Aaron is not aware of Dennett discussing the cognitive mechanisms of mathematical discovery).

AI vision methods

Methods used in machine vision, such as supervised learning (training an AI to label pictures of objects based on past samples) or having the AI build an internal 3D model of the environment, are likely not similar to how visual processing occurs in the human brain. Aaron believes that human visual experience likely involves a very large number of cognitive processes that are difficult to describe, and that for this reason it would be very difficult to test whether an AI subjectively experiences vision in the same way a human does. Some of the ideas about human vision were developed and demonstrated in a program called POPEYE in the mid-1970s, reported in Chapter 9 of *The Computer Revolution in Philosophy* (freely available online at http://www.cs.bham.ac.uk/research/projects/cogaff/crp/).

Overuse of statistical averaging

Aaron's impression is that many researchers in this area mainly interpret their data using statistical averages, which may preempt further investigation of notable variations in particular individuals. In particular, studying averages for groups of children at various ages rather than tracking individual changes is unlikely to yield deep insights into developmental mechanisms.

All Open Philanthropy Project conversations are available at
http://www.openphilanthropy.org/research/conversations