A conversation with Dr. Gary Drescher, July 18, 2016

Participants

- Dr. Gary Drescher – Artificial intelligence researcher, PhD, Massachusetts Institute of Technology
- 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 Dr. Drescher.

Summary

The Open Philanthropy Project spoke with Dr. Drescher 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 topics included potential types of algorithms that might be components of phenomenal consciousness (such as Dr. Drescher's "Cartesian Camcorder" and "qualia as gensyms" proposals), and the potential moral implications of potential variations in such cognitive algorithms.

Consciousness as a set of cognitive architectures

(Some context for the conversation.)

Luke's current guess is that phenomenal consciousness will eventually be describable as a (potentially extremely diverse) set of cognitive architectures, where a "cognitive architecture" is a (possibly simple or complex) set of algorithms interacting in a certain way — akin to how many scientists now think about "memory," "face recognition," "attention," and "model-based learning" (in humans, non-human animals, and software).

Luke thinks that one path toward clarifying the nature of consciousness may be to identify cognitive architectures that seem, from the designer's perspective, like they might produce many of the details of human phenomenal experience — i.e. architectures that seem like they might feel like human consciousness "from the inside." Such a set of algorithms would at least need to:

- Seem likely to produce something like typical human subjective experience under normal conditions,
- Provide explanations for illusions that occur in subjective experience,
- Provide explanations for confusions that humans experience about their own consciousness,
- Fit with known facts about ways in which self-reported subjective experience and behavior vary in response to brain damage, drug use, and other phenomena.

If an architecture of cognitive algorithms satisfying these conditions was identified, this alone might not suggest that such algorithms are very similar to the algorithms which actually implement human consciousness (let alone those which implement conscious experience in e.g. chimpanzees or pigs), but it could nevertheless help to narrow one's search for consciousness-instantiating algorithms within the space of all possible algorithms.

Luke sees Dr. Drescher's work on consciousness as being an instance of this sort of investigation that is especially understandable to those who find it more natural to manipulate concepts and metaphors from software engineering than to manipulate concepts and metaphors from the philosophy of mind.

Suggestions for algorithms potentially involved in consciousness

Dr. Drescher's intuitions about the types of cognitive algorithms that might produce human subjective experience, as well as his intuitions about animal and AI consciousness, are currently pre-theoretical, though he thinks it is possible that some theory might eventually lend support to his intuitions.

Cartesian Camcorder

Dr. Drescher has proposed a model in which some mental events (sensations, thoughts, etc.), unconscious at the moment that they occur, are "recorded" by an internal memory system that Dr. Drescher terms the "Cartesian Camcorder." A recorded event can be replayed and ''watched'' by other parts of the cognitive system (either immediately or later on), and it is this process that constitutes consciousness of the event. The mental event of watching the recording can itself be recorded and replayed, just as one can record a video of oneself watching a video.

The Cartesian Camcorder does not produce "literal" recordings (e.g. of raw sense data), but rather "smart" recordings that represent events at a higher level of abstraction, incorporating the agent's existing concepts and understanding of relations between the event, the agent, and other objects.

In this model, mental events will appear to be intrinsically conscious (despite not being so at the moment they first occur) in part because the act of ''looking at'' a mental event, via the Cartesian Camcorder, is what constitutes consciousness of the event. Thus, any mental event "examined" in this way, to check whether it is conscious, necessarily will be.

Qualia as gensyms

In the Lisp programming language, a "gensym" is a symbol that, as Drescher explains in *Good and Real*, "has no parts or properties, as far as the Lisp program can discern, except for its unique identity…" (p. 81). A Lisp program can compare two variables to determine whether or not the gensyms they point to are equivalent, but the program might have no other information about properties that distinguish the two variables.

Dr. Drescher hypothesizes that qualia might be implemented by cognitive

algorithms analogous to those which implement gensyms. These algorithms might "pass along" to the Cartesian Camcorder the fact that two sensations (e.g. light or no-light) are different, but *not* pass along any other information (e.g. about differences in the structure of the sensations, or anything about how they have been processed by the brain). In this model, the "ineffability" of qualia results from this layer of abstraction that prevents conscious access to the internal structure of a sensation.

It has been conventionally suggested that the ineffability of qualia is associated with the first-person "privileged access" of an individual to their own mental states. The "gensyms" model, conversely, suggests that the ineffability of qualia is due to an individual's *lack* of first-person access to the details of how a given sensation is computed, which an external observer could, in principle, access (e.g. the particular neuronal pattern that implements the sensation of red).

A generalization of Drescher's "qualia as gensyms" proposal might account for the partial ineffability of many different phenomena. For example, in the case of color qualia, a gensym-like algorithm might pass along comparisons of brightness, and comparisons of hue in the RGB field, in addition to simple discrimination of identity. If the algorithm passed information about the "distance" between colors in RGB space, that might provide a reason why orange feels "closer" to red than to blue in phenomenal experience, even though additional details about the structure of any particular color remains apparently ineffable. Perhaps a similar model could be generalized to sound and taste qualia, as well as to more complex sensations like pain. In each case, the Cartesian Camcorder would have access to the brute fact that two sensations are different, and might also have access to *some* additional information (e.g. that orange is "closer" to red than to blue), but would lack access to other information about the structure of, and relations between, differing sensations, and thus these sensations would seem to be partly or entirely ineffable, from the point of view of the Cartesian Camcorder.

"Qualia as gensyms" as a component of a broader set of cognitive algorithms

It seems somewhat plausible to Luke that a generalization of "qualia as gensyms" plus several other processes — Drescher's Cartesian Camcorder, some minimal form of self-modeling, etc. — could comprise a cognitive architecture that, from a designer's perspective, might seem as though it would instantiate something like human-like consciousness. Dr. Drescher agrees that this general approach might be productive, but worries that the relevant sciences (cognitive neuroscience, artificial intelligence, etc.) might not yet be advanced enough to develop a compelling model of human-like consciousness.

Situation-action vs. prediction-value systems

Dr. Drescher distinguishes between (1) "situation-action" systems, which implement compiled behavioral policies for responding to particular situations (e.g. as in thermostats, and perhaps insects), and (2) "prediction-value" systems, which combine representations of the potential effects of various actions in a situation

with valuations of those outcomes in order to select the action that maximizes expected value.

Dr. Dresher believes that pure situation-action systems likely do not have phenomenal consciousness. Luke's intuition is that, even if evolution has not yet produced pure situation-action machines that are conscious, it would in principle be possible to add some set of algorithms to a situation-action system that would instantiate phenomenal consciousness of those situation-action processes. Dr. Drescher agrees, noting that humans appear to have some situation-action programming (e.g., reflexive responses, such as sneezing), but also have a corresponding phenomenal impression of "what it is like" to sneeze.

Potential decoupling of prediction and valuation

It seems possible in principle that "prediction" (in the basic sense of world-modeling) could be decoupled from the "valuation" component of a prediction-value system (as might be occurring in, e.g., athymhormia patients). However, Dr. Drescher thinks that conscious experience seems to at least require the type of modeling system that could in principle be integrated into a full prediction-value system.

"Intrinsic" desirability or undesirability of conscious states

A common view among philosophers is that pleasure is intrinsically desirable, pain is intrinsically undesirable, and that humans act to pursue pleasure and avoid pain in recognition of this. Dr. Drescher suggests that, instead, humans are behaviorally hard-wired to tend to pursue or avoid certain sensations, and that the notions of "intrinsic desirability/undesirability" are reifications of those tendencies as observed in our own cognitive reactions and emotions.

The condition of pain asymbolia, in which patients do not exhibit a drive to avoid painful stimuli despite apparently receiving otherwise standard nociceptive signals (i.e. in which the "intrinsic undesirability" of the pain signal appears to be absent), may provide evidence for this view. From the computational perspective, pain asymbolia could be interpreted as a disruption of the usual coupling of pain sensations with avoidance behavior.

Effect of variations in cognitive algorithms on moral status

Even if a scientific explanation of the cognitive processes involved in consciousness were completed, people might still have different intuitions about the level or kind of moral status to assign to different cognitive systems. Any set of algorithms hypothesized to produce human consciousness would likely have the potential to vary along a number of dimensions, and some variations might warrant granting a cognitive system more or less moral consideration. Considering each feature of the set of cognitive algorithms, and determining how the moral weight one assigns to the system would change if the feature were altered or removed, could help reveal one's intuitions about the moral value of, e.g., AIs or animals, based on their relevant differences from humans.

Dr. Drescher's own views about moral status draw heavily on the notion of subjunctive reciprocity, and under such an approach, phenomenal consciousness (as usually conceived) might or might not be necessary for a being's moral status (see chapter 7 of *Good and Real*).

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