Testing for Synthetic Consciousness: The ACT, The Chip Test, The Unintegrated Chip Test, and the Extended Chip Test

Edwin Turner[1] and Susan Schneider[2]

[1] Dept. of Astrophysics, Princeton University
Ed.Turner@gmail.com
[2] Dept. of Philosophy, The University of Connecticut
Susan.Schneider@uconn.edu

Abstract. Despite the existence of several scientific and philosophical theories of the nature of consciousness, it is difficult to see how we can make progress on machine consciousness without some means of testing for consciousness in AIs. In short, we need to be able to "detect" conscious/subjective experience in a given AI system. In this paper, we present some behavior-based possibilities for testing for synthetic consciousness and discuss their potential limitations. The paper divides into several parts.

1 Consciousness Engineering

We briefly explain why the most sophisticated AIs (e.g., AGIs and "superintelligences") may not be conscious. Issues include architectures that do not feature conscious processing, the possibility that alternate substrates will not support consciousness, ethical concerns that militate against the use of sentient machines, and more. Further, AI developers may engineer consciousness in or out of different kinds of systems, depending upon social and economic factors. We then discuss several tests.

2 The ACT Test

One of the most compelling indications that normally functioning humans experience consciousness, although this is not often noted, is that nearly every adult can quickly and readily grasp concepts based on this quality of felt consciousness. Such ideas include scenarios like minds switching bodies (as in the film Freaky Friday); life after death (including reincarnation); and minds leaving "their" bodies (for example, astral projection or ghosts). Whether or not such scenarios have any reality, they would be exceedingly difficult to comprehend for an entity that had no conscious experience whatsoever. It would be like expecting someone who is completely deaf from birth to fully appreciate the experience of hearing a symphony.

An ACT would challenge an AI with a series of increasingly demanding natural language interactions to see how quickly and readily it can grasp and use concepts and

scenarios based on the internal experiences we associate with consciousness. At the most elementary level we might simply ask the machine if it conceives of itself as anything other than its physical self. At a more advanced level, we might see how it deals with ideas and scenarios such as those mentioned in the previous paragraph. At an advanced level, its ability to reason about and discuss philosophical questions such as "the hard problem of consciousness" would be evaluated. At the most demanding level, we might see if the machine invents and uses such a consciousness-based concept on its own, without relying on human ideas and inputs.

3 The Chip Test

As brain chips are further developed and refined as treatments for brain disorders, we anticipate their use in the areas of the brain underlying conscious experience (indeed, there are several relevant clinical trials for neural prosthetics going on now). We devise a test for synthetic consciousness based on neural prosthetics, and explain how this test can supplement ACT. We then identify two other kinds of "chip tests".

3.1 The Unintegrated Chip Test

We introduce a "chip test" to determine whether the integrated information theory (IIT) provides a necessary condition on AI consciousness. If neural prosthetics are successfully used in parts of the human brain that underlie consciousness and they lack suitable phi measures (we call these "unintegrated chips"), we have reason to believe that the sort of integrated information IIT proposes is not necessary for synthetic consciousness. (There are already important counterexamples to the sufficiency of IIT, such as Scott Aaronson's grid example.)

3.2 The Extended Chip Test (for testing the Extended Mind/Extended Consciousness Hypothesis)

This version of the Chip Test is a means of testing the extended mind hypothesis. The extended mind hypothesis says that the mind extends beyond the brain and body, into the world (Clark, Chalmers). As neural prosthetics are developed, we can pose a chip test for extended consciousness.

Let us suppose that someone passes the chip test. The following seems like a test that could be developed, at some future point: suppose the chip is outside of the person's head, but is still causally integrated in the same manner as before. If consciousness seems unaltered (by testing the patient in careful clinical settings) this seems to be a case for a strong version of the extended mind hypothesis: namely, consciousness extends beyond the head.

(For those familiar with the literature in philosophy on the metaphysics of personal identity, this test does not establish that the "extended" person, self or mind is the same person, self or mind as before).