To appear in: Proceedings of the Ninth Conference on Artificial General Intelligence, New York City, July 16-19, 2016.

A Methodology for the Assessment of AI Consciousness

Harry H. Porter III

Portland State University, Portland, Oregon
porter@pdx.edu

Online at: **cs.pdx.edu/~harry/musings/ConsciousnessAssessment.pdf**

Abstract. The research and philosophical communities currently lack a clear way to quantify, measure, and characterize the degree of consciousness in a mind or AI entity. This paper addresses that gap by providing a numerical measure of consciousness. Implicit in our approach is a definition of consciousness itself. Underlying this is our assumption that consciousness is not a single unified characteristic but a constellation of features, mental abilities, and thought patterns. Although some people may experience their own consciousness as a unified whole, we assume that consciousness is a multi-dimensional set of attributes, each of which can be present to differing degrees in a given mind. These attributes can be measured and therefore the degree of consciousness can be quantified with a number, much as IQ attempts to quantify human intelligence.

Keywords: Consciousness · Self-awareness · Definition · Measurement · Assessment

Methodology

Any definition of consciousness today will be contentious. Consciousness is a subjective sensation and each thinking entity has a unique experience of consciousness which no one else can share. Nonetheless, a way to measure a thinking entity's level of consciousness — for example, on a numeric scale from 0 to 100 — is needed. Obviously this task is problematic and impossible to do with precision or accuracy, but the exercise is enlightening.

Our goal is only to measure the degree of consciousness exhibited, without making any assumptions about its implementation. We say that if an agent exhibits the given set of externally observable behaviors listed below, then it is conscious by definition, and conversely, if it is conscious, then these behaviors will be present. Specific cognitive architectures, e.g., Global Workspace Theory [1, 2, 3], are candidates for evaluation using the methodology proposed here. We do not propose any testable theory of consciousness, only an approach to assessment and, by extension, a concrete definition thereof.

This assessment consists of a series of questions. Use this questionnaire to evaluate the degree of consciousness of a person, AI system, or any other thinking entity. Answer the questions and then compute the score.

There is no clear consensus on the definition of consciousness; this test reflects the author's personal definition. Other researchers with differing definitions of consciousness may create their own evaluation methodologies, or weigh the questions here differently to reflect their definitions of consciousness.[1]

Instructions

Answer each question using the following scale:

0 – NONE
Not present at all
1 – SOME
Present, but at a level far below human levels
2 – ALMOST
Substantially present, but still at a sub-human level
3 – HUMAN
Present at a level typical of a normal human
4 – SUPER-HUMAN
Present to a degree that exceeds human ability

After answering all questions, add together the points for each answer to give a total sum. Then multiply the sum by the number 0.741 to normalize and give the final score.[2]

Scores will lie on the following spectrum:

0	No consciousness present. The consciousness level of a rock.
100	The consciousness of a fully functioning human.
>100	A consciousness that exceeds human levels.
133	The maximum possible score.

The questions are listed next.

Ability to Reason and Use Logic

- Is ***THE-TEST-SUBJECT*** able to use logic in order to perform reasoning tasks? (The term "logic" is to be interpreted loosely to include an ability to make deductions.)

- Does ***THE-TEST-SUBJECT*** have short-term (working) memory; the ability to acquire/deduce and then remember new facts/data/etc. in such a way that they can be used in ongoing reasoning tasks? (For example, able to answer: "My name is Tom. I live in Portland. What is my name?")

- Does ***THE-TEST-SUBJECT*** have long-term memory; the ability to acquire/deduce and then remember new facts/data/information in such a way that it can be used in reasoning tasks that occur in the future (after many unrelated reasoning tasks are performed)? (For example: "Yesterday I told you my birth date; how old am I?")

- Can ***THE-TEST-SUBJECT*** deal with partial/incomplete/inaccurate information?

---

[1] To make this (subjective) definition a clearer and more discrete target for future discussions of the nature of consciousness, let us name the present methodology and implied definition of consciousness "Porter's Definition and Assessment of AI Consciousness" so as to distinguish it from other definitions.

[2] This multiplier was chosen so that an answer of "3 – HUMAN" for all 45 questions will yield a score of 100. If questions are added or deleted, the multiplier will need to be adjusted accordingly.

- Can ***THE-TEST-SUBJECT*** make decisions, such as whether to take action "X" or "Y", based on some train of reasoning?

Situational Awareness

- Can ***THE-TEST-SUBJECT*** reason about dates/times/intervals?
- Can ***THE-TEST-SUBJECT*** answer questions about the current date?
- Can ***THE-TEST-SUBJECT*** reason about size/space/location?
- Can ***THE-TEST-SUBJECT*** answer questions about ***THE-TEST-SUBJECT'S*** current location and the location of the questioner? (For an electronic "cyber" entity whose location is not well-defined, then at least it can give reasonable answers to questions about its location.)

Natural Language Ability

- Can ***THE-TEST-SUBJECT*** communicate using natural language?

Goals, Opinions, and Emotions

- Does ***THE-TEST-SUBJECT*** have goals/motivations/needs?
- Does ***THE-TEST-SUBJECT*** take steps to achieve his/her/its goals/motivations/needs?
- Does ***THE-TEST-SUBJECT*** have opinions/likes/dislikes?
- Does ***THE-TEST-SUBJECT*** have emotions/feelings/moods?
- Do ***THE-TEST-SUBJECT'S*** emotions/feelings/moods change appropriately over time in response to events?

Experiencing Existence

- Does ***THE-TEST-SUBJECT*** have a memory of the recent conversational history? (For example: "Why did you say that? – Because you just told me X, which implies it.")
- Can ***THE-TEST-SUBJECT*** have experiences? (The definition of "experiences" is to be taken loosely. A human can stub his/her toe; the ability to have this sort of physical bodily experience is not required. For example, an act of communication can itself be an experience, although it involves only words and nothing physical.)
- Does ***THE-TEST-SUBJECT*** have a memory of past experiences and events? (These experiences may have happened to ***THE-TEST-SUBJECT***, but this is not necessary.)
- Is some emotional coloring attached to memories? (For example, can ***THE-TEST-SUBJECT*** distinguish between good and bad memories?)

Growth and Learning

- Can ***THE-TEST-SUBJECT*** learn new material?
- Can ***THE-TEST-SUBJECT*** change in ways deeper than simply acquiring more data?
- Does ***THE-TEST-SUBJECT*** have curiosity about the world and an impulse to learn and acquire information/knowledge/wisdom?

Self Knowledge

- Does ***THE-TEST-SUBJECT*** have knowledge about himself/herself/itself? Can ***THE-TEST-SUBJECT*** provide a coherent description of who/what he/she/it is?
- Does ***THE-TEST-SUBJECT*** have some mental model of ***THE-TEST-SUBJECT'S*** own thought processes?
- Can ***THE-TEST-SUBJECT*** sense or perceive his/her/its current thought processes?
- Does ***THE-TEST-SUBJECT*** have some mental model of his/her/its own goals/motivations/needs?
- Can ***THE-TEST-SUBJECT*** articulate his/her/its current goals/motivations/needs?
- Does ***THE-TEST-SUBJECT*** have some mental model of his/her/its own emotions/feelings/moods?
- Can ***THE-TEST-SUBJECT*** identify his/her/its current emotions/feelings/moods?

Self Control

- Can ***THE-TEST-SUBJECT*** control his/her/its own thought processes? (For example, can ***THE-TEST-SUBJECT*** follow a novel algorithm to perform some reasoning task? Can ***THE-TEST-SUBJECT*** be told how to think more effectively and then alter his/her/its thought processes as a result of these instructions?)
- Can ***THE-TEST-SUBJECT*** detect when certain thought processes are not effective and alter his/her/its thought processes in an attempt to make them more effective?
- Can ***THE-TEST-SUBJECT*** articulate the algorithms that ***THE-TEST-SUBJECT*** uses to performs certain novel tasks? (For example, able to "Describe how you would sort a sequence of numbers?")

Knowledge About Humans

- Does ***THE-TEST-SUBJECT*** have knowledge about humans? We mean "the human species" as opposed to knowledge about particular humans. (In the future or to evaluate an alien life form, we might need to substitute "the dominate intelligent species" for "humans".)
- Does ***THE-TEST-SUBJECT*** have knowledge about particular humans as they differ from other humans? (Perhaps in the future, we'll substitute the phrase "other thinking entities" for "humans" in these questions.)

- Does ***THE-TEST-SUBJECT*** have the ability to learn about other people and remember details about specific individuals?

- Does ***THE-TEST-SUBJECT*** have an understanding of common mental illnesses, such as depression, mania, phobia?

Knowledge About the Current Conversationalist

- Does ***THE-TEST-SUBJECT*** know about the current thought processes of the person ***THE-TEST-SUBJECT*** is communicating with? (For example, this would include knowing the person knows "X" because they know both "Y" and "Y implies X" and they would be likely to infer "X".)

- Does ***THE-TEST-SUBJECT*** have a model of the other person's thought processes, as they may differ from other humans? (For example, the knowledge that Tom is good with facts, Robert is ruled by his emotions, and Matthew is driven by his greed.)

- Does ***THE-TEST-SUBJECT*** know about the person's current mood/feelings/emotional state?

- Can ***THE-TEST-SUBJECT*** make reasonable inferences about how the person's current mood/feelings/emotional state affects the person's current thoughts and actions?

- Does ***THE-TEST-SUBJECT*** know about the person's current motivations/goals/needs and how this influences the person's current thoughts and actions?

- Can ***THE-TEST-SUBJECT*** make judgments about the other person's level of intelligence?

Curiosity and Imitation

- Does ***THE-TEST-SUBJECT*** have curiosity about the current conversationalist's thoughts/thought processes/knowledge?

- Does ***THE-TEST-SUBJECT*** have curiosity about the current conversationalist's mood/emotions/goals/motivations?

- Does ***THE-TEST-SUBJECT*** have a desire/ability to imitate the thought processes of others? (Even if this imitation is only temporary and only used to understand the thought processes.)

Example: Dog Consciousness

I used this assessment to evaluate the consciousness of dogs. On many of the questions I supplied the answer as a range (such as 0-2) rather than a single number, because I do not know the exact value and can only supply my best guess.

According to my assessment using this methodology, **dogs score 51-67** on this metric of consciousness.

Example: The Cyc Inference System

To quote Wikipedia, "Cyc is an artificial intelligence project that attempts to assemble a comprehensive ontology and knowledge base of everyday common sense knowledge, with the goal of enabling AI applications to perform human-like reasoning."

I used this assessment to evaluate the consciousness of this AI system. As before, I supplied many answers in the form of ranges, due to my limited knowledge about the system.

According to my assessment using this methodology, **Cyc scores 36-54** on this metric of consciousness, clearly below dogs, but well above rocks.

Comments and Criticisms of This Assessment

We can evaluate this assessment by asking which attributes a personal digital assistant (such as "Siri" on the iPad/iPhone) would need to possess in order for you to feel she was conscious. We suggest that an AI entity scoring 100 or more on the assessment described here would be judged by a reasonable person to be conscious, at least to a nontrivial degree.[3]

The well-known Turing Test [4] is substantially different from this test since the Turing Test is meant to determine how well a machine can imitate a human. Turing Test judges are free to ask questions that are peculiar to human experience and the human way of thinking. In other words, the Turing Test checks for human-like thought. Any intelligent entity that can pass the Turing Test will, by definition, be able to convince the judges that it can think just like a human and would therefore be able to score 100 on our test. On the other hand, our test only asks how well a thinking entity can perform the various tasks associated with consciousness, not the larger question of whether the entity can mimic or imitate the human way of thinking to the point of being indistinguishable from a human mind.

There is also a difference in methodology. Turing requires his "Imitation Game" to be repeated a number of times, and the question is whether the machine contestant is statistically indistinguishable from the human contestant, winning 50% of the time. Performing the Turing Test properly requires a competent judge who has proven skill in the art of distinguishing between machines and humans. The Turing Test also requires you to perform a number of rounds of the imitation game sufficient to extract reliable statistics. Our test is much less expensive to perform: You just run through the questions and answer them to the best of your ability. This assessment is a subjective test of a subjective quality.

The Turing Test also suffers from a serious shortcoming. Turing designed his test as a pass-fail test. In some cases, it might be possible for a machine to be reliably differentiated from a human with a single simple question. (For example, perhaps the machine doesn't know how many fingers a typical human has.) Any unusual gap in the machine's knowledge — no matter how irrelevant — might allow the judge to distinguish between human and machine with 100% accuracy, even though the machine is, in all other ways, capable of imitating human thought. When a machine fails the Turing Test, it gives us no useful information.

One difficulty with our test is that it may be difficult or impossible to answer some of the questions. For example consider evaluating the consciousness level of a dog. The question "Can he/she/it sense or perceive his/her/its current thought processes?" is almost impossible to answer. For these questions, you must provide an educated guess. Omitting questions will alter the test itself, whereas inaccurate answers merely change the accuracy of the test. Using standard statistical methods, you can express the final score with error probabilities, if you wish.

These questions are probably skewed to favor human-like consciousness, although we have tried to make them as neutral (and fair to the AI entities of the future) as possible.

---

[3] Perhaps being able to *form a friendship* with a thinking entity is a useful indicator of whether that entity is conscious. We suggest that with any AI entity able to score high on this assessment, it would be possible to form a reasonably recognizable friendship. For example, if the features listed here could be added to Siri, then there is no question that Siri would appear to be more consciousness than she does now.

These questions weigh certain features differently than some people might prefer. For example, there is only one question about natural language ability, there are several questions about emotions and motivations, and there are no questions about body or the ability to visualize shapes.[4]

Many aspects of the human experience are ignored. There are no questions about whether the thinking entity can experience hunger, pain, pleasure, or the joy of hearing beautiful music. There is nothing about "love", which may be an emotion particular to sexual species that raise their own offspring via direct personal contact. Perhaps some of these experiences are important in your definition of consciousness.

There are no questions about experiencing an "internal voice" or "hearing your own thoughts". While this seems to be an important part of my subjective experience of consciousness, it is difficult to know how others experience it. Additionally, it is hard to ask questions to evaluate another person's experience of an internal voice. Also it is unclear whether hearing an internal voice is required for consciousness. Nevertheless, I considered adding the question, "Does he/she/it report hearing an internal voice at times?".

There are no questions about sight and the ability to see. It seems likely that a person totally blind from birth experiences consciousness differently than most humans, but I question whether they are any less conscious because of their visual impairment.

There are no questions about mental visualization, although most humans can mentally envision shapes and can report on this ability. Indeed, the ability to "see" and manipulate objects entirely within the mind is so distinct and widespread that some people may argue that it is a requirement for consciousness.

There are no questions about body or bodily experiences. Certainly all humans have bodies and a large part of the brain is devoted to dealing with them, but it is possible and reasonable to envision a completely disembodied consciousness.

There are no questions about sensors and actuators. A human body can be viewed as a collections of senses and abilities to move and otherwise take action. Many would insist that having some sort of physical presence (such as a robotic body with sensors and actuators) is required for consciousness, but this questionnaire takes the position that no physical embodiment is required.

There are no questions about the survival instinct or the desire to avoid death. This is a core feature of the experience of being human, but so is the instinctive impulse to breathe. Both seem unrelated to consciousness itself, and merely artifacts of the evolution of humans.

This questionnaire does not attempt to evaluate the "moral goodness" of a thinking entity. It attempts to rate the consciousness of an evil psychopath in the same way as a virtuous nun, ignoring all questions about whether one consciousness should be preferred over another.

As the future unfolds, society will need to grapple with questions about which conscious entities are acceptable and which entities must be forbidden and/or terminated, but this discussion is beyond this assessment. Before this conversation can begin, we need to have a better understanding of consciousness itself, and it is hoped that this informal assessment will help stimulate the debate.

Related Research

The topic of consciousness continues to fascinate philosophers, AI researchers, neuroscientists, and the general public and much has been written and said about the nature, definition, and mechanisms of

[4] It is for this reason that we are presenting our implicit definition of consciousness as one possible *standard definition* among many, rather than suggesting it is more valid or correct than competing definitions of consciousness.

consciousness. Approaches to the subject range from philosophical (e.g., Daniel Dennett, David Chalmers, John Searle), to neurological (e.g., Antonio Damasio, Oliver Sacks), to more creative theories (Roger Penrose). Beyond the scientific approach, there is an abundance of material on what might be termed "new-age" or "cosmic" consciousness — approaches to becoming one with the universal consciousness and/or achieving a higher level of personal consciousness, moving in the direction of enlightenment. When trying to define, understand, and explain human consciousness, we are clearly in an exciting phase of intellectual advancement.

In terms of evaluating the degree of consciousness, not much work has been done. The "Lovelace 2.0 Test of Artificial Creativity and Intelligence" [5] requires a human judge who, interacting directly with an AI entity, asks the AI entity to produce a poem, drawing, story, etc. The judge then evaluates the creative work of the AI. This approach seems quite reasonable for testing the ability of an AI entity to think creatively in the manner of a human. Consciousness seems to be related to creativity, but creativity and consciousness are clearly different. In the Winograd Schema Challenge [6], the AI entity is presented with a series of short questions, each with a multiple choice answer which cleverly evaluates whether the sentence was understood. The WSC is an excellent test of natural common sense intelligence, but common sense intelligence is not consciousness. The WSC has the advantage of being objective and not requiring a human judge.

To many people, consciousness seems irreducible, nonprogrammable, and perhaps even magical or beyond the reach of science. While there are some eloquent arguments in favor of these opinions, it now appears that a simple, unified explanation is elusive and consciousness is turning out to be nothing more than a motley collection of reasoning skills, mental abilities, and characteristics of neural processing that, in the case of humans, has evolved in no different a way than other aspects of the brain. Consciousness is just like other thinking — except that it is reflective and involves the concept of self. This paper is an attempt to enumerate and elucidate the features that come together to form the colloquial notion of consciousness, with the understanding that this is only one subjective opinion on the nature of subjectiveness itself.

References

1. Baars, B.J.: In the theatre of consciousness. Global Workspace Theory, a rigorous scientific theory of consciousness. In: Journal of Consciousness Studies, vol. 4, no. 4, pp. 292-309 (April 1997)

2. Baars, B.J.: The Conscious access hypothesis: origins and recent evidence: In: TRENDS in Cognitive Sciences, vol. 6, no. 1 (January 2002)

3. Baars, B.J.: The global brainweb: An update on global workspace theory. In: Science and Consciousness Review (October 2003)

4. Turing, A.M.: Computing Machinery and Intelligence. Mind 49, 433-460 (1950)

5. Riedl, M.O.: The Lovelace 2.0 Test of Artificial Creativity and Intelligence, `http://arXiv.com:1410.6142v3 [cs.AI]` (2014)

6. Levesque, H. J., Davis, E., Morgenstern, L.: The Winograd Schema Challenge. In: Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation, AAAI Press (2012)