Reviewing Tests for Machine Consciousness

Article *in* Journal of Consciousness Studies · June 2018

2 authors:

Aida Elamrani
Ecole Normale Supérieure de Paris

Roman Yampolskiy
University of Louisville

Reviewing Tests for Machine Consciousness

A. Raoult, R. Yampolskiy

November 28, 2015

Abstract

The accelerating advances in the fields of neuroscience, artificial intelligence, and robotics have been reviving interest and raising new philosophical, ethical or practical questions that depend on whether or not there may exist a scientific method of probing consciousness in machines. This paper provides an analytic review of the existing tests for machine consciousness proposed in the academic literature over the past decade and an overview of the diverse scientific communities involved in this enterprise. The tests put forward in their work typically fall under one of two grand categories: architecture (the presence of consciousness is inferred from the correct implementation of a relevant architecture) and behavior (the presence of consciousness is deduced by observing a specific behavior). Each category has its strengths and weaknesses. Architecture tests main advantage is that they could apparently test for qualia, a feature that is getting an increasing attention in recent years. Behavior tests are more synthetic and more practicable, but give a stronger role to ex post human interpretation of behavior. We show how some disciplines and places have affinities towards certain type of tests, and which tests are more influential according to scientometric indicators.

1 Introduction

The accelerating advances in the fields of neurobiology, artificial intelligence, and robotics have been reviving interest and raising new philosophical, ethical or practical questions that depend on whether or not there may exist a scientific method of probing consciousness in machines. This increasing interest can be seen in the academic world, with the apparition of specialized journals (such as International Journal of Machine Consciousness) and dedicated conferences (ECAL[1], BICA[2], FLI[3], ...) and the high frequency of publication of scholarly

[1] In 2013 the European Conference on Artificial Life hosted the first international workshop on artificial consciousness.

[2] In 2014, the 5th annual meeting of the Biologically Inspired Cognitive Architectures Society held a symposium entitled "Toward Self-Conscious Robotics".

[3] In 2015, the Future of Life Institute held their first conference with "Creating human-level AI: how and when?" being one of the four themes addressed.

books [55][9][27], as well as in popular culture (the movies “Her” in 2013, “Transcendence” in 2014, “Ex Machina” in 2015).
Defining what consciousness is and what it might mean for a machine to be conscious is a task that draws on centuries of philosophical inquiry ([?],[54],[?]). One of the most problematic issues raised by machine consciousness is known in philosophy as the “other minds problem”: how can I know that anyone beside myself feels something[?]?
Machine consciousness tests are a specific subquestion of this general problem. Since the famous Turing Test of 1950[53] [4], dozens of tests, more or less practical, general and unique, have been proposed to offer a procedure that, if followed, would indicate whether or not a machine should be deemed ”conscious”. This paper provides an analytic review of the existing tests for machine consciousness discussed in the academic literature over the past decade and an overview of the diverse scientific communities involved in this enterprise.
We first expose the method used to gather the relevant literature. Second, we give a brief description of each test, explaining the theories upon which they are built. Third, we analyze them by answering five simple questions: 1- how do they operate? 2- what features of consciousness are being tested? 3- who are the authors? 4- where do they come from? 5- how much impact do they have? Finally we offer some thoughts on how these different aspects imply a typology, and a relation between the background of the researchers and their approach.

2 Methodology

This article presents a review of the tests for machine consciousness appearing in the scientific literature in the past 10 years (2004-2014). The data collection followed two steps.
The first step consisted in performing a search on Google Scholar for different combinations of the keywords ”test” and ”consciousness” since 2004[5]. We collected a total of 359 publications. We have then filtered out studies targeting animals or humans to save only the ones concerned with machines or minds in the broad sense, which resulted in a little less than 40 publications. Among these, only 18 suggested an evaluation method of machine consciousness.
The second step we followed was to search the references. We searched the bibliographies of each of the 18 results acquired in the first step for similar publications since 2004. This provided us with 14 additionnal publications.
At this stage we had 32 publications, including only papers published in peer-reviewed journals, conference proceedings, books, and a single working paper from arxiv [52], from 30 different authors. However, after a thorough examination of the cited references, we noticed that 4 publications dated before 2004

[4]Which cautiously set aside the sensitive question of consciousness to concentrate more pragmatically on intelligence[?].

[5]The request was: ”consciousness testing” OR ”consciousness test” OR ”testing consciousness” OR ”test for consciousness”. Interestingly enough, looking for ”qualia testing” OR ”qualia test” OR ”testing qualia” OR ”test for qualia” returned only 4 results for the same time frame, all of which were already returned by the former request.

seemed to be too important to skip in our review; they were all cited at least once in our initial publication set by someone else than their original author. Finally, our review includes a total of 36 publications and 35 authors.
Because we wanted to compare several characteristics of these numerous tests and put them in perspective with the machine consciousness community, we built a relational database[6] and meticulously encoded our results following a process that we describe in 4 giving us the ability to extensively manipulate the data and test several hypothesis.

3 Description - A chronological review of contemporarily relevant tests

In what follows, we are presenting all the propositions we found currently discussed in the peer-reviewed literature.

1950: The Turing Test Incontestably the most influential experiment of this kind, with more than 7000 citations on Google Scholar, the Turing Test was originally designed as a test for machine intelligence, pioneering the field of A.I. half a decade before this term was coined[35]. Today, some authors suggest that the Turing Test could suffice to indicate that a machine is conscious [25][41].
Procedure: Isolate the machine in a room, a human competitor in another room, and a human judge in a third one. The judge can communicate by text through teletype with both competitors, but without seeing them. The A.I. is as intelligent as a human if the human judge is unable to distinguish it from the human competitor. [53]

1970: The Mirror Test Initially designed in 1970 by Gallup [16] to experimentally assess self-recognition in primates, the experiment consisted in three stages:
Procedure:

1. familiarizing the chimpanzee with its reflection by placing a mirror in its cage.
2. anesthetizing the chimpanzee to paint a red mark on its ear.
3. observe how the chimpanzee reacts: if the chimpanzee interacts with the mark on its ear, self-recognition is acquired.

Since its publication, the Mirror Test has been used on other animals besides chimpanzees [39][38][48] as well as in infants [5][8]. The Mirror Test has then been transposed to robots multiple times to validate their self-awareness [45] [20] [14] which is presented as a decisive feature of consciousness[46].

[6] That will be made publicly available with the final version of this paper

1991: The Total Turing Test Among the answers to Searle's famous "Chinese Room" thought experiment [42], which was meant to expose how computers are incapable of genuine understanding, the notion of embodiment -providing the machine with a body- was put forward by several authors as necessary to instill intentionality within the machine[26][15]. Following this line of thought, Harnard proposes in 1991 an updated version of the Turing test that he names the "Total Turing Test" [**?**].
Procedure: The A.I. needs to be embodied in a robot, it cannot consists solely of a program running on a computer. Furthermore, this robot has to be able to do "in the real world of objects and people"[**?**] anything a human can do.

2001: The Lovelace Test Dissatisfied with Turing's answers to Lovelace's concern that the machine never "originates" anything but merely follows a set of instructions[**?**], Bello, Ferrucci and Bringsjord design a new test for thinking machines aimed at assessing their creativity: is the machine capable of forming completely novel thoughts and unpredictable associations? This issue was identified by Cariani [**?**] as "creative emergence", the ability for a computer to autonomously develop new *informational primitives*, as opposed to "combinatoric emergence" which can only result in combinations of previously established primitives.
Procedure: As stated in [12]: an artificial agent, A, designed by a human, H, passes the Lovelace Test if and only if:

1. A outputs o;
2. As outputting o is not the result of a fluke hardware error, but rather the result of processes A can repeat;
3. H (or someone who knows what H knows, and has Hs resources) cannot explain how A produced o by appeal to As architecture, knowledge-base, and core functions

In the Lovelace Test the machine's speech abilities do not bear as much importance as in the Turing Test. In fact, there is no restriction whatsoever on the domain in which the machine should express its creativity (literature, music, mathematics,...). However the candidate is required to be able to communicate its production for a human judge in order to pass the test. At any rate, in the event of a verbal evaluation, *how* the speech was conceived would be the crucial criteria rather than the meaning it conveys.

2003: ADM Axioms Aleksander, together with Dunmall, introduces an axiomatic definition of consciousness in [3] aimed at describing the necessary mechanisms an "agent need[s] to possess for it to have a sensation of its own presence and affordances" using "what is known of the neurology of the brain". Later, he publishes an updated account of this theory with Morton in [2]. We call this approach the "ADM Axioms", after the initials of each contributor. **Procedure:**The machine's architecture must observe the following axioms, as defined in [2]

1. Presence: I feel that I am an entity in the world that is outside of me.
2. Imagination: I can recall previous sensory experience as a more or less degraded version of that experience. Driven by language I can imagine experiences I never had.
3. Attention: I am selectively conscious of the world outside of me and can select sensory events I wish to imagine.
4. Volition: I can imagine the results of taking actions and select an action I wish to take.
5. Emotion: I evaluate events and the expected results of actions according to criteria usually called emotions.

Although the authors do not provide a generic method to verify these axioms within any given architecture, Aleksander designed a cognitive architecture, called the Kernel Architecture[4] in accordance with ADM.

2004: Phi (Φ) Based on the integrated information theory of consciousness (IIT) introduced by Tononi in 2004 [50], Φ is the result of a sophisticated ad-hoc equation measuring the quantity of integrated information in a system. The underlying theory explains consciousness as a graded phenomenon corresponding to the amount of information integration achieved by a physical system (Φ), associating human-level consciousness with a high value of Φ. This theory describes subjective experience as a *conceptual structure* and specifies the five axioms of experience that can account for the accomplishment of these structures by any physical system:

1. Existence: the system must have causal power, "its mechanisms must be able to 'make a difference' to the probability of its past-future states" [52].

2. Composition: An experience, or conceptual structure, can be described as a set of elementary *concepts*.

3. Information: The difference between experiencing red or blue is informational. "A system of mechanisms in a state specifies a differentiated conceptual structure: each structure is the specific way it is (a specific composition of concepts), differing in its specific way from other possible ones."[52]

4. Integration: A conceptual structure is unified and irreducible. "Irreducibility can be measured as integrated information [Φ], which quantifies to what extent the conceptual structure specified by a systems mechanisms changes if the system is partitioned (cut or reduced) along its minimum partition (the one that makes the least difference)."

5. Exclusion: A conscious experience is carried through a *maximally irreducible* conceptual structure called a *complex*.

The initial formulation of Φ has been refined several times since its creation, until very recently [51][49][37], but the general idea of the method has not changed. While Φ is theoretically applicable to any physical system, it is currently not practically computable [37]. However it does offer another specific angle on how consciousness in machines might be tested. Koch and Tononi expect that modern computers would "likely not form a large complex of high Φ^{max}, but break down into many mini complexes of low Φ^{max}" [52].
Procedure: As explained in [52], to calculate Φ for a computer "one should analyze its real physical components, identify elements, say transistors, define their cause-effect repertoires, find concepts, complexes, and determine the spatio-temporal scale at which Φ reaches a maximum."

2005: The Ordinal Consciousness Scale Gamez proposes a pragmatic method to evaluate the consciousness of any given system, S: he develops a heuristic scale to calculate the similarity between S and the human brain.[17][18]
Procedure: As given in [18]: The position, omc_{pos}, of an actual machine on a scale with omc_{tot} possible positions is found by calculating its total weighting, and looking for this value in the complete list of possible machines. To facilitate some kind of comparison between different versions of the scale, omc_{pos} is normalized to a value between 0 and 1 to give the final OMC rating which gives a rating of 1 for human brains and a rating close to zero for the last system on the list.

1. Rate

   R1 Approximately the same speed as the human brain 1.0

   R2 Ten times faster or slower than the human brain 0.55

   R3 Over a hundred times faster or slower than the human brain 0.1

2. Size

   S1 Approximately the same size as the human brain 1.0

   S2 A thousand times larger or smaller than the human brain 0.55

   S3 More than a million times larger or smaller than the human brain 0.1

3. Function Implementation

   F1 Produced by a biological structure of neurons 1.0

   F2 Produced by a non-biological structure of neurons 0.55

   F3 Produced using mathematical algorithms, computer code or some other method 0.1

4. Function of neurons

    FN1 Produced by a biological structure of molecules, atoms and ions 1.0

    FN2 Produced by a non-biological structure of molecules, atoms and ions (silicon chemistry, for example) 0.7

    FN3 Produced by a non-biological structure of neurons 0.4

5. FN4 Produced using mathematical algorithms, computer code or some other method 0.1

6. Function of molecules, atoms and ions

    FMAI1 Produced by real subatomic phenomena, such as protons, neutrons and electrons 1.0

    FMAI2 Produced by a non-biological structure of neurons 0.55

    FMAI3 Produced using mathematical algorithms, computer code or some other method 0.1

7. Time Slicing

    All functions are dynamically changing and co-present at any point in time 1.0

    TS2 Some functions are dynamically changing and co-present at any point in time 0.55

    TS3 A single function is dynamically changing and present at any point in time 0.1

8. Analogue / digital

    AD1 Analogue system 1.0

    AD2 Mixture of analogue and digital 0.55

    AD3 Digital system 0.1

Gamez has estimated the consciousness of Demarse's neurally controlled animat [?] at 0.990, the robot Lucy [?] at 0.505, Franklin's naval dispatching system IDA [?] at 0.245 and Block's "Chinese Nation" [?] between 0.745 and 0.02604 depending on the implementation.

2007: The Haikonen Mirror Test Haikonen [?] proposes an enhanced Mirror Test specifically designed for cognitive robots by requiring the presence of a somatosensory grounding system to guarantee the self-recognition validity. According to this theory, self-consciousness will emerge from a such a genuinely self-aware robot provided that its architecture involves a unified attention system. In his own terms:

> ”true self-awareness should involve a self-concept or a self-model that contains a body image and a mental self-image which are grounded to the material self via a somatosensory system. [...] Self-consciousness would arise, when the self-percepts and self-models become reportable via unified system attention.”[**?**]

Procedure:

1. Check the machine's architecture for the presence of a somatosensory system and unified attention system.
2. Follow the mirror test.

2007: The Higher Order Syntactic Thought (HOST) In [40] Rolls argues that grounded higher order syntax is a sufficient condition for consciousness and can lead to the emergence of qualia. According to this theory, current computers are not conscious because their syntax is not grounded semantically in the external world.
Procedure: This approach displaces the way consciousness in a machine can be tested. The question is now : can we prove "higher order linguistic thought supervisory correction process to correct first-order one-off linguistic thoughts with symbols grounded in the real world" ?

2007: Insight Learning In [**?**], Kohler observes how a chimpanzee manages to stack wooden boxes or assemble small wooden sticks into a longer one to reach a banana hung to the ceiling. This ability to solve problems by making new relations between previously acquired knowledge rather than through trial and errors is called *insight learning*. In [34], the authors claim that consciousness is the "ability to simulate behavior mentally", and suggest that this behavior could stand as a proof for machine consciousness.
Procedure: Design a test adapted to the machine to demonstrate its ability to make insight learning.

2007: The Haikonen Consciousness Test In [20], Haikonen proposes that a "flow of inner speech *with grounded meanings*" while maybe not necessary for consciousness is nonetheless a strong indicator of its presence. Therefore he proposes a "hallmark-based" test around these properties of human thinking.
Procedure: Show that the machine has inner speech or mental imagery and is aware of it, for instance by transforming inner speech into text. Furthermore, the construction of the machine has to be analyzed to certify that this mental content is not the product of a "preprogrammed string of words and imagery".

2008: Ψ-D Filter In the 2000s, several computational models of the mind inspired by psychoanalysis are introduced: synapse state theory of mental life [**?**], artificial recognition system-psychoanalysis [**?**], and machine psychodynamics [**?**]. In [1], the authors propose a trial based on three fundamental concepts of machine psychodynamics, defined *functionally*: *pleasure, tension and ambivalence.*

> A tension relates to the degree to which a part of a robots body deviates from its state of resting or to which a drive (such as fear, anxiety, excitation, boredom, or expected pain) deviates from its homeostatic equilibrium. The pleasure is a function of tension dynamics. When the tension abruptly drops, the pleasure volume rises and then slowly decays.(...)
> The term ambivalence refers to a perpetual struggle of ideas in the robots working memory. So, unlike a conventional robot that in the case of contradictory ideas is expected to possibly quickly decide which of them to implement, a psychodynamic robot may change its mind in unpredictable moments.

The authors introduce the term "proto-intentionality", which designates both the causal power in the machine behavior and it having goal-oriented actions, as a meaningful feature of consciousness and define the *psychodynamic criterion* as "the ability to achieve a state defined as pleasurable by deliberately plunging oneself into a state defined as unpleasant". They design an experiment to verify the observance of this criterion in robots behavior.
Procedure: Place the robot in a habitat with decaying vital resources, separated from the rest of the environment by an unsafe zone. The robot has no information on neither the size of the unsafe zone, nor on what lies beyond. If the robot gets *pleasure* from venturing in the unsafe zone, it satisfies the *psychodynamic criterion.*

2008: Quantum Opting New quantum theories of consciousness are regularly proposed [24][29][47]. This evaluation of machine consciousness is based on one of them. In [30], the authors "argue that quantum theory is a theory of alternatives in that, for a given experiment in which an observable quantity (such as the energy) of a system is measured, it specifies the alternatives that are open to the system" and "that from a given set of alternatives specified by quantum theory, individual quantum systems spontaneously opt for one alternative rather than another, and that this opting is a primitive form of consciousness, not reducible to any more fundamental category". They call this capacity to opt for a specific alternative a *primitive consciousness* and derive from it the suggestion that the architecture of a Turing machine can never endorse this aptitude. This argument parallels the concerns expressed by Bringsjord in [12] of machine creativity in the Lovelacean sense, e.g. as the capacity to *originate*, but here applied at the quantum level.
Procedure: Prove that a system is capable of quantum opting.

2008: ConsScale Introduced in [6] and improved in [7] ConsScale describes progressive levels of artificial consciousness each of them specified as a set of criteria defining abstract architectural components and cognitive skills. They

range from level 0:*disembodied*, biologically equivalent to an amino acid, to level 11:*super-conscious*, above human-like consciousness. An agent can only be assigned a given level if it complies to the requirements set for all the inferior levels. They further propose two normalized scores to serve"as a quantitative measure of cognitive power associated with artificial consciousness": 1-cumulative levels score (CLS), 2-consscale quantitative score (CQS).
Procedure: Locate the artificial agent on the scale by observing its behavior and architecture.

2008: The Koch and Tononi Test As an alternative to measuring Φ, but still with the intent of probing information integration Koch and Tononi propose a simple test that requires the machine to "differentiate the scene's key features from the immense range of other possible scenes" [31], they offer some examples in [33] and [32].
Procedure: Present the machine with a picture containing a planted anomaly, for instance that of a computer with a flower pot instead of a keyboard standing in front of the screen, and ask as formulated in [32]: "what's wrong with this picture?"

2009: The P-Consciousness Test Initially formulated in [21], improved in [22] following [43] this test is aimed at proving evidence of p-consciousness, as defined by Block [11]. As described by Hales:

> "The P-Consciousness Test can be viewed as an empirically viable variant of the Total Turing Test and the Lovelace Test obtained by choosing a single very specialised behaviour: scientific behaviour"[23].

The authors propose that the test accounts for ADM axioms. Hales gives an illustration of the test in [21] and [23].
Procedure: The machine has to understand a law of nature.

2010: Robot Philosopher In 2010 Sloman suggests that if a robot is able to ponder the world to the point of reaching modern philosophers's questionning, then it is conscious. Even though the test was not intended as a serious proposition [44], it has was carefully considered and it raised a lot of attention in the machine consciousness community, eventually leading to a special issue of International Journal of Machine Consciousness [**?**].
Procedure: Observe the machine's thoughts until they become comparable to modern philosophers questions.

2012: Cross-examination Proposed in [**?**], this test is very similar to the Turing Test but even easier for the machine. Haikonen suggest to let the machine convince us of its consciousness. **Procedure:**If the machine has the correct architecture, simply ask whether it is sentient. This seems to be a change of heart from the author, who previously insisted on how "we can easily be fooled" [20] by the ability of a program to mimic behavior.

2012: The Qualia Total Turing Test This test is another improvement of the Total Turing Test. The questions asked to the machine must involve qualia. [**?**]
If one is unable to distingush between a human answering question about qualia and a machine doing the same, then the test is passed. **Procedure:** The procedure is a variation of the Total Turing Test but the interrogation must involve qualia (for example, through wine tasting).

2012: Jazz Improvisation Chella and Manzotti suggest a new variant of the Turing test, based on a robot (embodiement necessary) that would be able to improvise jazz.
Procedure: The procedure is a variation of the Total Turing Test where the judge has to appraise a jazz improvisation.

2014: Human-Like Conscious Access In [19], Goerztel introduces what he calls "the six key factors of human-like consciousness" which aim at unifying global workspace theory [**?**][**?**] and the integrated information theory of consciousness and are presented as necessary manifestations of experience:

> 1. Dynamical representation: the entity X should correspond to a distributed, dynamic pattern of activity spanning a portion of the system (a probabilistically invariant subspace of the systems state space). Note that X may also correspond to a localized representation, e.g. a concept neuron in the human brain
> 2. Focusing of energetic resources: the entity X should be the subject of a high degree of energetic attentional focusing
> 3. Focusing of informational resources: X should also be the subject of a high degree of informational attentional focusing
> 4. Global Workspace dynamics: X should be the subject of GWT style broadcasting throughout the various portions of the systems active knowledge store, including those portions with medium or low degrees of current activity. The GW functional hub doing the broadcasting is the focus of energetic and informational energy
> 5. Integrated Information: the information observable in the system, and associated with X, should display a high level of information integration
> 6. Correlation of attentional focus with self-modeling: X should be associated with the systems self-model, via associations that may have a high or medium level of conscious access, but not generally a low level.[19]

As a measure of consciousness based on these six factors, he proposes two indicators of human-like consciousness: a ratio and a degree.
Procedure: First, quantify the six factors. Then, "to formalize the degree to which a system S gives human-like conscious access to an entity X, (...) quantify conscious access as a weighted combination of these factors, with the weighting being state of consciousness dependent."[19]

4 Analysis

We first analyze how the test operate and which features of consciousness they address, then who proposes these tests and where they come from. Finally we consider their impact with respects to scientometrical criteria.

4.1 How do the tests operate?

To compare these tests we have narrowed a set of key characteristics evaluated for each test by answering a series of simple yes/no questions. The results of this evaluation are transcribed in table 1.
Because the tests presented themselves in under a great variety of forms, some questions had to be adapted. Below, we give for each question every formulation it has taken:

1. Explicit. Is the test formulated explicitly to determine if a machine is conscious to some extent? Otherwise, it means that a necessary or sufficient characteristic for machine consciousness is proposed (as in HOST [40]).

2. Architecture. Are the machine's architecture/functions decisive factors to pass the test? Can a machine that passes the test be dismissed because of its architecture/functions?

3. Behavior. Is the machine's behavior (or functional equivalent) a decisive factor to pass the test?

4. Model Testing. Is the test designed to suit a specific model of consciousness? Which model?

5. Stronger than Turing. Does a machine that passes this test necessarily passes the Turing Test?

6. Verbal. Would it be possible to decide of the results with a typed interview of the machine?

7. Communication. Is the machine somehow required to be able to communicate with humans?

8. Human Design. Is human-level interpretation necessary to design an implementation of the test?

9. Human Results. Is human-level interpretation necessary to decide for the outcome?

10. Measurement. Does it result in a numerical measure?

11. Application. Is there an example of application of this test?

12. Subject. Can this test be used to decide if anything (A)/a computer (C)/a robot (R)/a human (H)/an animal (N) is conscious?

Name	Explicit	Architecture	Behavior	Model Testing	Stronger than Turing	Verbal	Communication	Human Design	Human Results	Numerical Measure	Application	Anything	Computers	Robots	Humans	Animals
Turing Test	✓		✓			✓			✓				✓	✓	✓	
Mirror Test	✓		✓						✓				✓	✓	✓	✓
Total Turing Test	✓		✓		✓		✓		✓					✓		
Lovelace Test	✓	✓					✓	✓	✓				✓	✓		
5 Axioms	✓	✓		✓				✓	✓				✓	✓	✓	✓
Phi	✓	✓		✓				✓		✓		✓	✓	✓	✓	✓
OMC Scale	✓	✓						✓		✓			✓	✓	✓	
Haikonen Mirror Test	✓	✓				✓		✓	✓				✓	✓		
HOST		✓		✓			✓	✓				✓	✓	✓	✓	✓
Insight Learning	✓		✓					✓	✓				✓	✓		
Haikonen Consciousness Test	✓	✓	✓			✓			✓				✓	✓		
Psy-D Filter	✓		✓	✓				✓					✓	✓		
ConsScale	✓	✓	✓	✓				✓	✓	✓			✓	✓	✓	✓
KTT	✓		✓					✓					✓	✓	✓	
Quantum Opting	✓	✓						✓				✓	✓	✓	✓	✓
PCT	✓		✓				✓	✓					✓	✓	✓	
Robot Philosopher	✓		✓		✓	✓	✓		✓					✓		
Cross-Examination	✓	✓	✓			✓	✓	✓	✓				✓	✓		
Q3T	✓		✓		✓	✓	✓	✓	✓				✓	✓	✓	
Jazz Improvisation	✓	✓	✓				✓		✓				✓	✓		
Human-Like Conscious Access	✓	✓		✓				✓		✓		✓	✓	✓	✓	✓

Table 1: Tests characterization

First, the most striking observation that appears from this characterization is that *human evaluation is always required at some point*: a human is systematically needed, either to design the test or to read the results. This means that all the tests we reviewed draw on Turing's proposal, where a human judge decides whether or not a machine is indistinguishable from a human - the main difference between each test being the context, constraints or elements of the judgment. In any case, we have found no method that would allow for the construction of a testing device (e.g. a box where one could place a machine, with a green led expected to flash when the machine is conscious). The proposal that comes nearer to the possibility of such a device is quantum opting, but this test is also the most speculative one as it suffers from a lack of knowledge in this area of scientific study, making any practical implementation impossible.

Second, it is worth noting that only two tests require an analysis of both the architecture and the behavior of the machine. Aside from these two exceptions, there is a clear separation between, on the one hand, architecture-based tests which generally work by identifying necessary mechanisms or interacting modules of consciousness, and on the other hand, behavior-based tests which typically require the machine to achieve a specific action. Interestingly, another dichotomy is generally accepted for computational models of consciousness with on one side, functional models relying on behavior, and on the other side phenomenological approaches [2] but this distinction does not stand in our study. We found that architecture can be defined functionally (Haikonen mirror test, ConsScale,...), and behavior can be used to assess phenomenological properties (Q3T, jazz improvisation).

By combining the two preceding remarks, a third observation quickly follows: tests that consider architecture will typically require human evaluation during the implementation process to appraise the mechanisms and decide whether or not a system complies to a set of given prerequisites for consciousness, while tests concerned with behavior tend to require human interpretation of the results. To illustrate this notion, let us consider the case of an architecture-based test, Φ: a human is required to inspect the machinery and determine which components (for instance, transistors) are relevant to the system before being able to proceed with the real test which consists in a calculation; for a behavior test like the Mirror Test, a human is needed to decide how to interpret the subject's reactions. Therefore we can conceptualize consciousness tests as variations and precisions built on the Turing test triangular model (the judge, the machine, the human benchmark), where an architecture-type test is interested in how the design is up to the standards of the human benchmark, while the behavior-type test is interested in how the behavior of the machine compares to the one exhibited by the human benchmark (hence closer to the original Turing Test).

4.2 Which features of consciousness are being tested?

We have established for each test and each reference a list of the features of consciousness that they claim to address, with no judgment from our part on the accuracy of the claim. Most of the time our observations were straightforward, recording the exact words as they appeared in the text. We selected 14 features which occurred repeatedly without interpretation, alphabetically: attention, awareness, creativity, dynamism, emotions, grounding, imagination, intelligence, intentionality, language, qualia, perception, self, volition. But in some cases we made the choice to adapt the original vocabulary either by 1- resorting to a synonym to match the most widespread usage and allow for a better comparison between texts, 2- breaking down a complex concept into smaller equivalent units. The equivalences we made are given below.

1. Creativity: novelty [23]; imagination defined as the ability to "recall or fabricate"[3]
2. Dynamism: parallel processing "the processing of functions can be carried out in parallel with all of them operating simultaneously on dedicated hardware" [18]; distributed, dynamic pattern of activity spanning a portion of the system [19]
3. Grounding: "situated" [13]
4. Qualia: "subjective experience" [13] "first-person", "phenomenology" [2]
5. Perception: functional depiction of the external world [3]; weaker than "qualia"
6. Self: "existence" [52] ;"presence" [2];
7. Information integration: awareness + self

We found that there is a strong correlation between the test orientation (architecture vs behavior) and the tested features. This is expressed by the graphs 1 to 3 (1, 2, 3). Each node represents a tested feature, and two nodes are linked by an edge if they appear in the same test. Edges are weighted, so when two features are more often tested by the same test, the edge appears larger. Nodes size is also weighted: the more a feature appears in the tests, the larger it will be on the graph. The graph layout was computed with Gephi using the ForceAtlas2 algorithm, a procedure in which the nodes repulse and the edges attract[28].

Graph 1 shows the main preoccupation of the reviewed tests is the self, while dynamism and intentionality are marginal.

Architecture-based tests are represented on graph 2. They stand out as being more specific tests, mainly preoccupied with perception, self and qualia.

Behavior-based tests are represented on graph 3. The graph is denser because behavior-based test are concerned with more features. Intelligence becomes

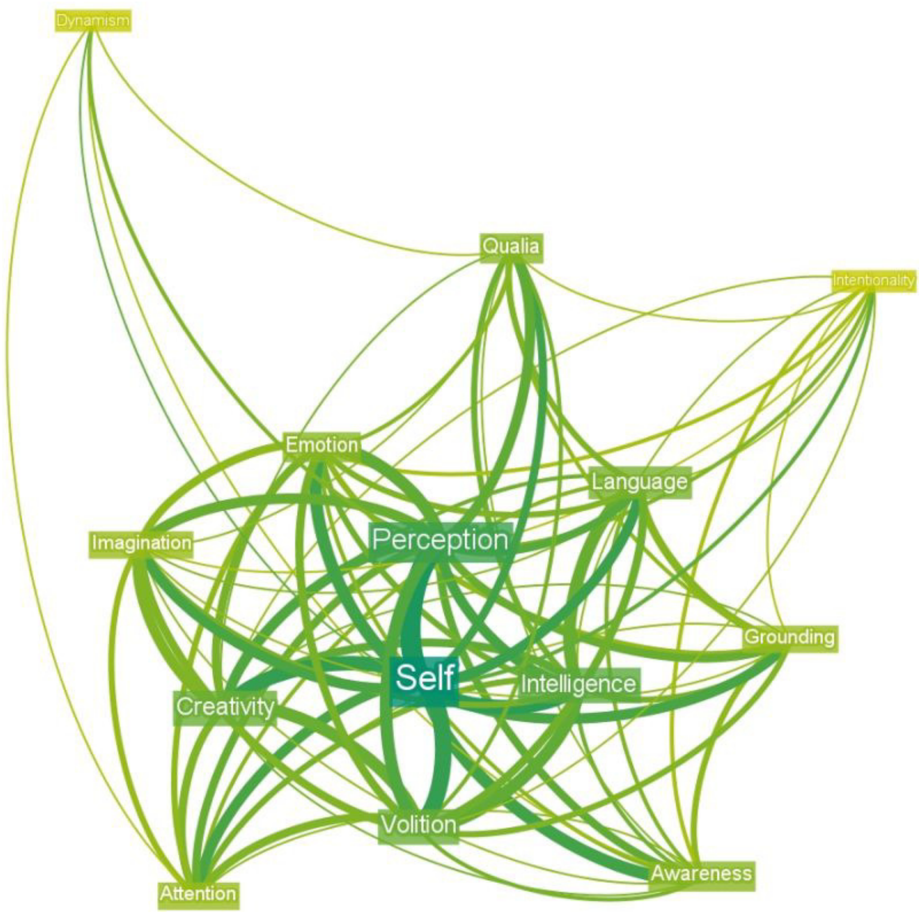

Figure 1: Graph of features co-occurrence, all tests included

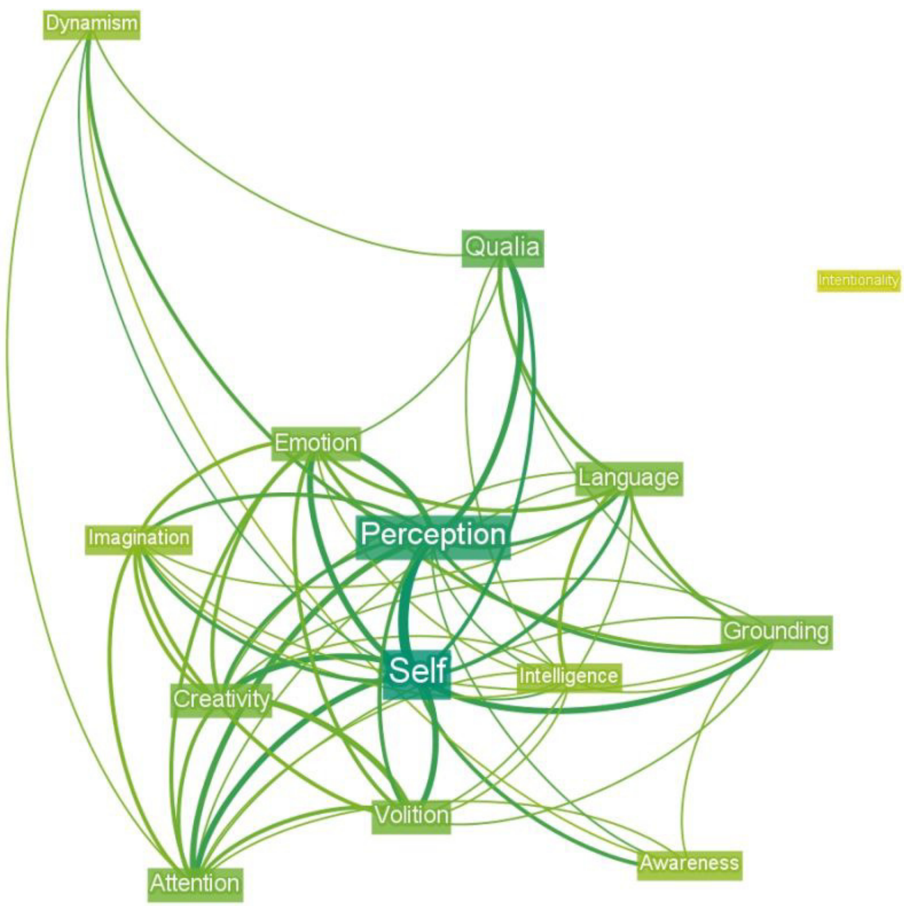

Figure 2: Graph of features co-occurrence for architecture tests

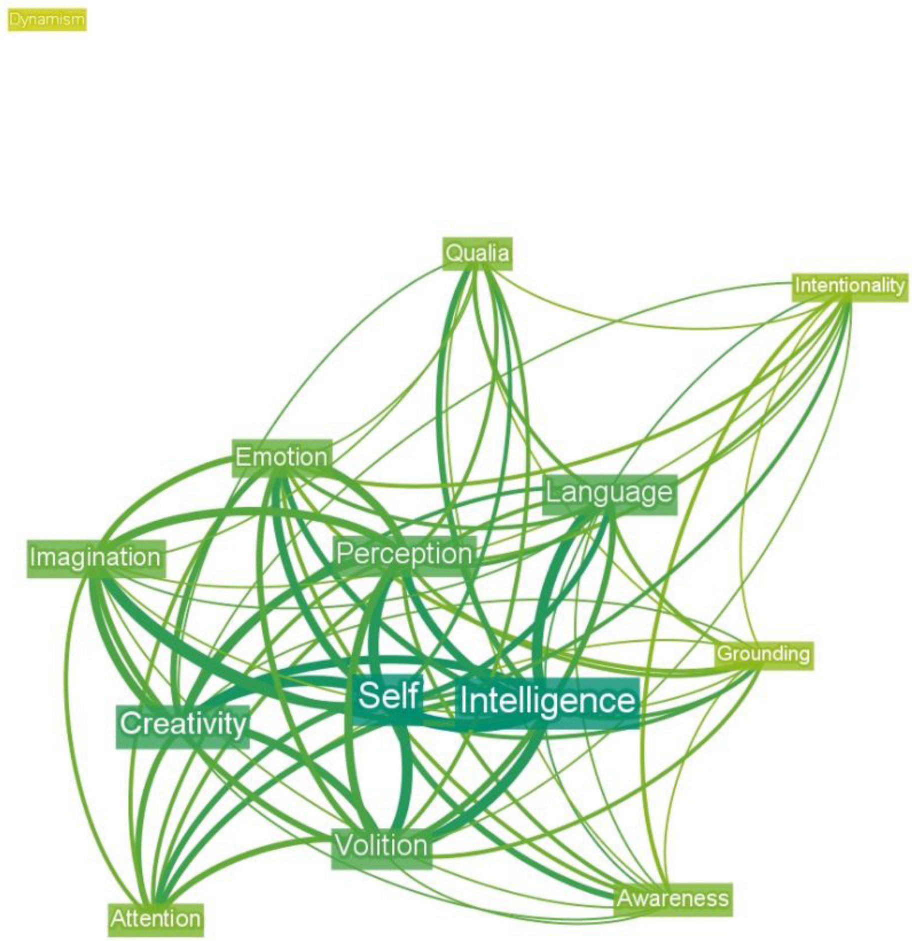

Figure 3: Graph of features co-occurrence for behavior tests

almost as important as the self (this corresponds to Turing variants), and creativity. This difference highlights the synthetic aspects of behavior tests which imply a large role in human interpretation but allow for a more global definition of consciousness.

4.3 Who proposes the tests?

There is a high diversity of profiles among researchers publishing on this matter, which reflects in part the diversity of consciousness studies in general: neuroscience, humanities, physics play a major role. Neuroscience's role is obvious since human brain is commonly considered as "the gold standard for consciousness" [34][**?**]. Humanities are implicated because of the questioning common to the philosophy of the mind; finally, physics also because machines and brain alike ultimately break down to elementary particles.
To this known group, we of course have to add robotics and computer science that enter the machine consciousness question through the machine angle instead of the consciousness angle. Using Scopus[7], we were able to retrieve the total number of each author's publications, classified following ASJC classification [8], the results are shown in table 2 and with more details in table 3. It is to be noted that a single publication can be classified under several subject areas. We found that computer science is the most widespread discipline in this community, with 33 active authors, while, still regarding the machine expertise, engineering comes 4th (26 authors). Humanities come 3rd, with 26 authors, 19 of which published in philosophy. Authors tied to biology are slightly more prolific in biochemistry, genetics and molecular biology (25 of them - 18 in "Ecology, Evolution, Behavior and Systematics") than neuroscience (23 - 15 of which in cognitive neuroscience). Mathematics play an important role in this community, with 13 researchers having published in "Modeling and Simulation" and 12 in "Applied Mathematics".
There is no obvious influence of the major disciplines (computer science, neuroscience, engineering, humanities) on the type of test, however it is striking that those major disciplines are home to more original profile than their peers. On the minor disciplines, there are predictable outcomes of disciplinary differences with the major ones. Physicists tend to have a lighter emphasis on speech and communication abilities (55% of tests) than neuroscientists (78% of tests) and mathematicians are close to physicists (60%). On the other hand, model testing approaches are more common among physicists (55%) and theoretical computer scientists (45%) than among neuroscientists (28%) and philosophers (28%). In between the minor disciplines, there are also differences. No psychologist test for intentionality, while 18% of theoretical computer scientists and 13% of mathematicians test for it.

[7]Scopus is Elsevier's peer-reviewed literature online database.

[8]Elsevier "All Science. Journals Classification", publicly available at http://www.info.sciverse.com/documents/files/scopus-training/resourcelibrary/xls/title_list.xls

	Subject Area	**# of authors**
1	Computer Science	33
2	Mathematics	29
3	Arts and Humanities	26
4	Engineering	26
5	Biochemistry, Genetics and Molecular Biology	25
6	Social Sciences	25
7	Psychology	24
8	Neuroscience	23
9	Medicine	20
10	Agricultural and Biological Sciences	19
11	Physics and Astronomy	13
12	Decision Sciences	10
13	Multidisciplinary	9
14	Chemical Engineering	9
15	Pharmacology, Toxicology and Pharmaceutics	8
16	Environmental Science	8
17	Business, Management and Accounting	8
18	Chemistry	7
19	Earth and Planetary Sciences	5
20	Materials Science	4
21	Energy	3
22	Health Professions	3
23	Nursing	2
24	Immunology and Microbiology	2
25	Economics, Econometrics and Finance	1
26	Veterinary	1
27	Dentistry	1

Table 2: Number of authors with at least one publication by general subject area

	Subject Area	auth		Subject Area	auth
1	Artificial Intelligence	26	33	Bioengineering	7
2	Control and Systems Engineering	23	34	Physiology	7
3	Software	23	35	Neurology (clinical)	6
4	Theoretical Computer Science	21	36	Ecology	6
5	Computer Science Applications	21	37	Pharmacology	6
6	Philosophy	19	38	Arts and Humanities (miscellaneous)	6
7	Electrical and Electronic Engineering	19	39	Sociology and Political Science	6
8	Information Systems	18	40	History and Philosophy of Science	6
9	Ecology, Evolution, Behavior and Systematics	18	41	Education	6
10	Computational Theory and Mathematics	17	42	Religious Studies	6
11	Human-Computer Interaction	17	43	Developmental and Educational Psychology	6
12	Experimental and Cognitive Psychology	15	44	Information Systems and Management	6
13	Cognitive Neuroscience	15	45	Statistics and Probability	6
14	Computer Vision and Pattern Recognition	15	46	Computational Mathematics	6
15	Hardware and Architecture	14	47	Biochemistry	5
16	Modeling and Simulation	13	48	Condensed Matter Physics	5
17	Psychology (miscellaneous)	12	49	Control and Optimization	5
18	Applied Mathematics	12	50	Computational Mechanics	5
19	Computer Networks and Communications	11	51	Biophysics	5
20	Signal Processing	11	52	Statistical and Nonlinear Physics	5
21	Language and Linguistics	11	53	Logic	4
22	Molecular Biology	10	54	Atomic and Molecular Physics, and Optics	4
23	Neuropsychology and Physiological Psychology	10	55	Management of Technology and Innovation	4
24	Biotechnology	10	56	Psychiatry and Mental Health	4
25	Linguistics and Language	10	57	Drug Discovery	4
26	Multidisciplinary	9	58	Neurology	4
27	Cultural Studies	8	59	Communication	4
28	Cellular and Molecular Neuroscience	8	60	Literature and Literary Theory	4
29	Computer Graphics and Computer-Aided Design	8	61	Medicine (miscellaneous)	4
30	Neuroscience (miscellaneous)	7	62	Mechanical Engineering	4
31	Behavioral Neuroscience	7	63	Biomedical Engineering	4
32	Genetics	7			

Table 3: Number of authors with at least one publication by detailed subject area, showing only subject areas with 4 or more results

Country	Currently Affiliated Authors	City	Currently Affiliated Authors
United States	10	?	1
		Arlington	1
		Iowa City	2
		Madison	2
		San Diego	1
		Seattle	1
		Springfield	1
		Yorktown Heights	1
United Kingdom	8	Birmingham	1
		Edinburgh	1
		London	3
		Oxford	1
		Southampton	1
		Sussex	1
Japan	5	Sapporo	2
		Tokyo	2
		Wako	1
Spain	3	Madrid	3
Italy	2	Milan	1
		Palermo	1
Australia	1	Parkville	1
Poland	1	Gdynia	1
India	1	Patna	1

Table 4: MC testing research around the world

4.4 Where do the tests come from?

With Scopus, we have also been able to find the authors current academic affiliations, see table 4. Europe, and most notably U.K., is the most active community, followed by the US and Japan.
The geographic differences are more striking that the disciplinary ones. Europe has a high priority for creativity (56%), emotion (67%) and perception (78%) while U.S. researchers are less interested by these features (30%, 10% and 40% respectively). Architecture-based approaches are common to the U.S. (60%) and Europe (56%) but scarce in the rest of the world (20%).
On all other aspects the distribution is homogeneous.

4.5 How impactful are the tests?

The last aspect of consciousness testing we review are scientometrics. We look on Google Scholar for the citation number of the most cited paper advocating the test, as well as the Scopus H-index of the author with the highest score.
With no surprise, Turing test is by far the most cited one we review, but the H-index recorded for Alan Turing by Scopus is abnormaly low (11), as Scopus

Name	Citations (Google Scholar)	Authors h-index (Scopus)
Turing Test	7452	11
Phi	585	80
Total Turing Test	175	20
5 Axioms	108	9
Haikonen Consciousness Test	63	5
KTT	58	80
Lovelace Test	46	10
HOST	45	102
Mirror Test	32	5
Robot Philosopher	28	1
OMC Scale	19	6
ConsScale	18	5
Haikonen Mirror Test	13	5
Cross-Examination	11	5
Quantum Opting	6	12
PCT	6	1
Psy-D Filter	5	6
Jazz Improvisation	1	11
Insight Learning	1	5
Human-Like Conscious Access	0	9
Q3T	0	*NULL*

Table 5: Tests ordered by total citations and authors' maximal h-index

is not well adapted for older research (in fact, the Turing Test referenced in Scopus is not the original 1950 publication but a recent re-edition).
Three other tests are cited more than 100 times : Φ, the Total Turing Test and the ADM axioms.
Most of the heavily cited tests emanate from heavily cited authors, with the exception of the Haikonen Consciousness Test and the ADM axioms, which denotes that for those two tests, machine consciousness is the primarily field of influence of the researchers, whereas for Φ, KTT and HOST, we have influential authors with one contribution to machine consciousness that is as influential as the rest of their work.

5 Conclusion

The tests that are reviewed in this paper typically fall under one of two categories: architecture-based tests (the presence of consciousness is inferred from the correct implementation of a relevant architecture) and behavior-based tests (the presence of consciousness is deduced from a behavior indistinguishable from

that of a human being). We showed how some tested features, authors discipline and affiliation have affinities toward certain type of tests, and which tests are most influential according to scientometric indicators.
Overall, this review highlights the apparent strength and weaknesses of the two approaches. Architecture tests seem the most theoretically promising but the least practicable, their main advantage being that they can claim to verify physical attributes or correlates of consciousness. Furthermore, their *ex ante* analysis of the machine's design is the most straightforward solution to spot and disqualify potentially undesirable candidates as look-up tables, which are commonly considered as mindless artifacts[10][9]. Behavior tests are more synthetic, they cover more features of consciousness, and they seem easier to put in practice. However, they give a strong role to *ex post* human interpretation of the observed behavior, which may pose reproducibility and transparency issues (how do I know if I would have personally judged the computed made jazz improvisation to be indistinguishable from the human made one?).
In the end, the architectural approach takes the human brain as the ultimate benchmark, mainly drawing from neuroscience, whereas the behavioral one is concerned with the human mind, and is tied more closely to psychology. In this respect, we could have expected the former approach to be limited to human-like consciousness and the latter one to accept a broader understanding of consciousness including new forms of artificial mentality, but authors from both sides seem equally dispersed on this opinion. Their areas of expertise do not seem to have an influence on the matter either. In this review we found the main problem to be the complex nature of consciousness, as illustrated by the multitude of different features evaluated by each test. It would seem wiser and less prone to debate to test for specific aspects of consciousness, to avoid the fact that no scientific consensus exists on this definition.

References

[1] Syed I Ahson, Senior Member, and Andrzej Buller. Toward Machines that Can Daydream. 2008.

[2] I Aleksander and H Morton. Why axiomatic models of being conscious? *J Consc Studies*, 14(7):15–27, 2007.

[3] Igor Aleksander and Barry Dunmall. Axioms and Tests for the Presence of Minimal Consciousness in Agents. *Journal of Consciousness Studies*, 10(4-5):7–18, 2003.

[4] Igor Aleksander and Helen Morton. Computational studies of consciousness. *Progress in brain research*, 168:77–93, January 2008.

[5] B. Amsterdam. Mirror self-image reactions before age two. *Developmental Psychobiology*, 5(4):297–305, 1972.

[9] Even thought this view has been disputed recently, see [36].

[6] Raúl Arrabales and A Ledezma. Criteria for consciousness in artificial intelligent agents. *ALAMAS&ALAg Workshop at*, 2008.

[7] Raúl Arrabales, Agapito Ledezma, and Araceli Sanchis. ConsScale: A Pragmatic Scale for Measuring the Level of Consciousness in Artificial Agents. *Journal of Consciousness Studies*, 17(3-4):131–164, 2010.

[8] Lorraine E. Bahrick and John S. Watson. Detection of intermodal proprioceptivevisual contingency as a potential basis of self-perception in infancy. *Developmental Psychology*, 21(6):963–973, 1985.

[9] R Blackford and D Broderick. *Intelligence Unbound: The Future of Uploaded and Machine Minds*, volume 9781118736. 2014.

[10] Ned Block. Psychologism and Behaviorism. *The Philosophical Review*, 90(1):5–43 CR – Copyright © 1981 Duke University P, January 1981.

[11] Ned Block. On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18(02):227, February 1995.

[12] Selmer Bringsjord, Paul Bello, and David Ferrucci. Creativity , the Turing Test , and the (Better) Lovelace Test. *Minds and Machines*, 11:3–27, 2001.

[13] A Chella and R Manzotti. Jazz and machine consciousness: Towards a new turing test. In *AISB/IACAP World Congress 2012: Revisiting Turing and His Test: Comprehensiveness, Qualia, and the Real World, Part of Alan Turing Year 2012*, pages 49–53, 2012.

[14] P Fitzpatrick, A Arsenio, and E R Torres-Jara. Reinforcing robot perception of multi-modal events through repetition and redundancy and repetition and redundancy. *Interaction Studies*, 7(2):171–196, 2006.

[15] J. A. Fodor. Searle on what only brains can do. *Behavioral and Brain Sciences*, 3(03):431, February 1980.

[16] G G Gallup Jr. Chimpanzees: Self-recognition. *Science*, 167(3914):86–87, 1970.

[17] D. Gamez and D. Gamez. An ordinal probability scale for synthetic phenomenology. *Proceedings of the 2005 AISB Workshop on Next-Generation Approaches to Machine Consciousness*, pages 85–94, 2005.

[18] David Gamez. *The Development and Analysis of Conscious Machines*. PhD thesis, 2008.

[19] Ben Goertzel. Characterizing Human-like Consciousness: An Integrative Approach. *Procedia Computer Science*, 41:152–157, 2014.

[20] P.O. Haikonen. *Robot Brains Circuits and Systems for Conscious Machines*.

[21] Colin Hales. An Empirical Framework for Objective Testing for P-Consciousness in an Artificial Agent. *The Open Artificial Intelligence Journal*, 3(1):1–15, February 2009.

[22] Colin Hales. The Well-Tested Young Scientist. *International Journal of Machine Consciousness*, 02(01):35–39, June 2010.

[23] Colin G. Hales. *The Revolutions Of Scientific Structure*, volume 3. World Scientific Publishing Company, 2014.

[24] Stuart Hameroff and Roger Penrose. Conscious events as orchestrated space-time selections. *Journal of Consciousness Studies*, 3(1):36–53, 1996.

[25] Stevan Harnard. Can a machine be conscious? How? *Journal of Consciousness Studies*, 10(4-5):67–75, 2003.

[26] John Haugeland. Programs, causal powers, and intentionality. *Behavioral and Brain Sciences*, 3(03):432, February 1980.

[27] P Hingston. *Believable bots: Can computers play like people?* 2013.

[28] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS ONE*, 9(6):1–12, 2014.

[29] E R John. A field theory of consciousness. *Consciousness and cognition*, 10(2):184–213, June 2001.

[30] David E Klemm and William H Klink. Quantum Physics and Beyond. 43(2):307–327, 2008.

[31] Christof Koch and Giulio Tononi. Can Machines Be Conscious. *IEEE Spectrum*, (june):55–59, 2008.

[32] Christof Koch and Giulio Tononi. A Test for Consciousness. *Scientific American*, (June):44–47, 2011.

[33] Christof Koch and Giulio Tononi. Testing for Consciousness in Machines. *Scientific American Mind*, 22(4):16–17, August 2011.

[34] C Marcarelli and J L McKinstry. Testing for machine consciousness using insight learning. In *AAAI Fall Symposium - Technical Report*, volume FS-07-01, pages 102–107, 2007.

[35] John McCarthy, Marvin L Minsky, Nathaniel Rochester, and Claude E Shannon. A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. *AI Magazine*, 27(4):12, 2006.

[36] Drew McDermott. On the Claim that a Table-Lookup Program Could Pass the Turing Test. *Minds and Machines*, 24(2):143–188, March 2014.

[37] Masafumi Oizumi, Larissa Albantakis, and Giulio Tononi. From the phenomenology to the mechanisms of consciousness: Integrated Information Theory 3.0. *PLoS computational biology*, 10(5):e1003588, May 2014.

[38] Joshua M Plotnik, Frans B M de Waal, and Diana Reiss. Self-recognition in an Asian elephant. *Proceedings of the National Academy of Sciences of the United States of America*, 103(45):17053–7, November 2006.

[39] D Reiss and L Marino. Mirror self-recognition in the bottlenose dolphin: a case of cognitive convergence. *Proceedings of the National Academy of Sciences of the United States of America*, 98(10):5937–42, May 2001.

[40] Edmund T Rolls. A computational neuroscience approach to consciousness. *Neural networks : the official journal of the International Neural Network Society*, 20(9):962–82, November 2007.

[41] Rafal Rzepka and Kenji Araki. Consciousness of Crowds The Internet as Knowledge Source of Human Conscious Behavior and Machine Self-Understanding. In *AAAI Fall Symposium - Technical Report*, pages 127–128, 2007.

[42] J. R. Searle. Minds, brains, and programs, 1980.

[43] Aaron Sloman. An Alternative To Working on Machine Consciousness. *International Journal of Machine Consciousness*, 02(01):1–18, June 2010.

[44] Aaron Sloman. Machine Consciousness: Response To Commentaries. *International Journal of Machine Consciousness*, 02(01):75–116, June 2010.

[45] Junichi Takeno. Experiments and examination of mirror image cognition using a small robot. pages 493–498, 2005.

[46] Junichi Takeno. *Creation of a Conscious Robot: Mirror Image Cognition and Self-Awareness*. Pan Stanford Publishing, 2012.

[47] Max Tegmark. Consciousness as a state of matter. *Chaos, Solitons & Fractals*, 76:238–270, July 2015.

[48] Koji Toda and Michael L Platt. Animal cognition: monkeys pass the mirror test. *Current biology : CB*, 25(2):R64–6, January 2015.

[49] G Tononi. Integrated information theory of consciousness : an updated account. *Arch Ital Biol*, 150:56–90, 2012.

[50] Giulio Tononi. An information integration theory of consciousness. *BMC neuroscience*, 5:42, November 2004.

[51] Giulio Tononi. Consciousness as Integrated Information :. (December):216–242, 2008.

[52] Giulio Tononi and Christof Koch. Consciousness: Here, There but Not Everywhere. *arXiv*, 2014.

[53] AM Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950.

[54] Gottfried Wilhelm Freiherr von Leibniz. *The monadology and other philosophical writings.* Oxford University Press, H. Milford, 1898.

[55] Roman V. Yampolskiy. *Artificial Superintelligence: A Futuristic Approach.* Chapman and Hall/CRC, 2015.