A novel approach for identifying a human-like self-conscious behavior

Gianpiero Negri

Abstract

In this paper, a possible extension of Turing test [1] will be presented, which is intended to overcome the limits highlighted by several researchers and scientists in the last seventy years.
The main problem related to the execution in Turing test is substantially dealing with the trouble in identification of a human-like intelligence based on a pure evaluation of external behavior of a machine.
In this work first of all a description of classical Turing test will be done. After that, some of the main exceptions or oppositions to the Turing test ability to detect "intelligent machines" will be presented.
The Lovelace test will be presented as well, as possible alternative to Turing Test, and some considerations on its scope and effectiveness will be made.
Furthermore, some references to Penrose and Hofstadter ideas will be recalled, highlighting the strongest troubles in defining and detecting a human-like intelligence, intended as "self-consciousness".
Finally, the new approach will be explained, introducing the new test intended to overcome the troubles highlighted on Turing test execution, based on a model of the self-consciousness obtained by means of the hypersets theory. An example will be presented as well, in order to clarify the proposed approach and its goal.

1. Introduction: The Turing test

The description of Turing test is included in [1].
A human interrogator is located in a room, with two keyboards and monitors on two desks A and B; in a second room on the other side of a separating wall there are a computer C (for example connected to desk A) and another person, so that the first person (let's call him the "interrogator") can use both keyboards to chat with the other person and the computer.
The conversation is absolutely free, and the interrogator can choose to ask both computer and human being the same questions, or to talk about the same topics.
The goal of the test from the interrogator perspective is to distinguish between the human being and the computer.
If the human interrogator is not able to detect which is human, which is machine, the computer C passes the test.
Hence, it can be said that a computer able to demonstrate a human-like behavior, so that a human being is not able to distinguish it from another human being, is considered to have or show a certain form of "intelligence".
So this test is supposed to be a sort of "intelligence proof" for a computer.
Several scientists and researchers have shown a significant opposition to the above argumentation.
One of the most known arguments constraining the Turing test intent conclusion (that is, a machine can be considered "intelligent" if it behaves as a human for the interrogator) is presented by Searle [2], with his well-known example of the "Chinese room".
A human being, located in a closed room, is requested to translate in English some messages written in Chinese and coming from an outer room through a small window. The person in the "Chinese room" is not able to understand the Chinese, but on the other hand he can use a complete collection of manual or documents, including any possible Chinese symbol and a large collection of common phrases, so that he's supposed to be able to translate any kind of common message from Chinese to English.
Searle question was: is that translator able to understand Chinese? The obvious answer is "no, for sure".

2. Original and 2.0 Lovelace test

First of all, the definition of "intelligent agent" will be introduced [9]:

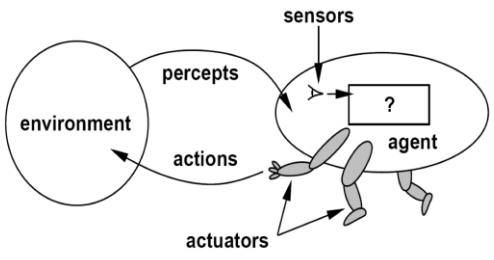

Figure 1: Intelligent agent schematic architecture

The overall architecture of an intelligent agent is shown in Figure 1: according to this architecture, agent take percepts from the environment, process them in some way that prescribe actions, perform these actions, take in new percepts, and continue in the cycle.
Hereafter, the terms "machine" and "intelligent agent" will be used in the same meaning.
With above definition in mind, the original Lovelace Test [7], [9], can be introduced: it attempts to formalize the notion of origination and surprise.
An artificial agent *a*, designed by *h*, passes the Lovelace Test if and only if:

• *a* outputs *o*,
• *a*'s outputting *o* is the result of processes a can repeat and not a fluke hardware error, and
• *h* (or someone who knows what *h* knows and has *h*'s resources) cannot explain how *a* produced *o*.

One critique of the original Lovelace Test is that it is unbeatable; any entity *h* with resources to build *a* in the first place and with sufficient time also has the ability to explain *o*. Even learning systems cannot beat the test because one can deduce the data necessary to produce *o*.
In [7] an updated Lovelace Test is proposed, as alternative to the Turing test.
The new Lovelace Test proposed asks an artificial agent to create a wide range of types of creative artifacts (e.g., paintings, poetry, stories, architectural designs, etc.) that meet requirements given by a human evaluator. A limited form of the new test asks that an artificial agent operate only be able to generate a single type of artifact. The Lovelace 2.0 Test is a test of the creative ability of a computational system, but the creation of certain types of artifacts, such as stories, require a wide repertoire of human-level intelligent capabilities.
The above test is designed to challenge the premise that a computational system can originate a creative artifact. We believe that a certain subset of creative acts necessitates human-level intelligence, thus rendering both a test of creativity and also a test of intelligence.
The Lovelace 2.0 Test is as follows: artificial agent *a* passes the Lovelace Test if and only if:

• *a* creates an artifact *o* of type *t*,
• *o* conforms to a set of constraints *C* where $c_i \in C$ is any criterion expressible in natural language,
• a human evaluator *h*, having chosen *t* and *C*, is satisfied that *o* is a valid instance of *t* and meets *C*, and
• a human referee *r* determines the combination of *t* and *C* to not be impossible.
According to the author of [7], the constraints set C makes the test Google-proof and resistant to Chinese Room arguments. An evaluator is allowed to impose as many constraints as he or she deems necessary to ensure that the system produces a novel and surprising artifact.
The reported example is: "create a story in which a boy falls in love with a girl, aliens abduct the boy, and the girl saves the world with the help of a talking cat."
While *C* does not necessarily need to be expressed in natural language, the set of possible constraints must be equivalent to the set of all concepts that can be expressed by a human mind.
The evaluation of the test is simple: a human evaluator is allowed to choose *t* and *C* and determine whether the resultant artifact is an example of the given type and whether it satisfactorily meets all the constraints.
The human referee *r* is necessary to prevent the situation where the judge presents the agent with a combination of *t* and *C* that are impossible to meet even by humans. The referee should be an expert on *t* who can veto judge inputs based on his or her expert opinion on what is known about *t*.
While original Lovelace test is considered as limited by the author of [7], according to its definition, Lovelace Test 2.0 can be in the same way considered only a "reinforcement" of Turing Test, in the sense that it has been built to be much harder to be passed by a machine, considering the need to satisfy the constraints and, in the same time, to produce a satisfactory output for (each possible) human interrogator.
It has been noted that, just as for the Turing Test, the Lovelace 2.0 Test has as scope for the artificial agent's creator (programmer) to use smart techniques that essentially deceive the judges into thinking the agent is being creative, whereas it just is able to perform an even more sophisticated sort of "perfect deception".
Moreover, according to the author of [7]: "the test provides no threshold at which one can declare an artificial agent to be intelligent. However, the test provides a means of quantitative comparing artificial agents. Creativity is not unique to human intelligence, but it is one of the hallmarks of human intelligence. Many forms of creativity necessitate intelligence". So it seems reasonable to conclude that both Turing and Lovelace (original and 2.0 version) are not able, and probably, have not the scope, to verify if a machine or artificial agent is "intelligent" or "self-conscious" as a human being is supposed to be.

3. Self-consciousness: a possible definition

As expressed in previous section, the Turing and Lovelace tests provides an un-decidable result, in the sense that there is no evidence of presence of self-consciousness for a machine able to pass them; the only thing that can be said for sure, as per definition, is that the machine is performing a sort of "perfect deception" [3].
At this point, let's consider two possible alternatives:

1 – Hypothesis of the algorithmic consciousness: as described in Hofstadter [4], considering the possibility that self-consciousness can be an "emergent property" of a system, in the future it will become possible to define an algorithm complex enough to produce the self-consciousness as consequence of its own complexity. In other terms, the intelligence will emerge in a system if the system itself will become enough complex.

2 – As described in Penrose [5], if the self-consciousness has some "constituting elements" with a non-algorithmic nature, something completely different from an "algorithm" must be developed (or created) in order to obtain a self-conscious behavior from a machine.

Note that, in both cases, by means of a Turing test it's not possible to evaluate if the machine has reached a self-consciousness or not.
Let's start this new approach description with the consideration that a self-conscious system can be defined as a being able to perceive itself, through different levels.
Basically, at least 3 levels of "ability" can be defined:

1. Execute actions: "I take the piece of paper from the table in front of me"
2. Reflect on itself doing actions: "I reflect on myself taking the piece of paper"
3. Reflect on its own ability to reflect on itself doing actions: "I'm thinking about myself while thinking about myself taking the piece of paper"

(and so on…)

A self-conscious system shall present, at least, all those 3 abilities or capabilities.
Considering 1. and 2. , it can be noted that, nowadays, there are machines able to execute actions and, in a certain sense, to think about or monitor themselves doing actions (an example could be an industrial machine performing some tasks as moving or manipulating objects and monitoring its own behavior with a diagnosis sensor architecture or something similar).
Considering 3. instead, it should be clear that it's very hard, or better, impossible, to find today a machine showing a similar kind of ability.
The meaning of 3., in fact, is that the system or the machine should be able to perceive itself as an unitary and definite entity, performing the "abstraction jump" on a level allowing the system to separate itself from the rest of the universe, including in the "universe" its own actions and its own thought about the actions it's able to perform. This property is something related to the so-defined "apperception" introduced by Immanuel Kant and Gottfried Leibnitz.
The whole set of ability of a being to perceive itself while perceiving itself doing each possible action ("I know that I know that I do") can be a good definition of "self-consciousness" of the being itself.
As consequence, a system having the capability expressed in 3. (let's call it S3), just because it has this capability, is able to bypass itself acting or thinking about itself acting.
For example, S3 should be able to formulate a judgment on a its own action while executing it about the "sense of executing the action".
Let's suppose that a machine executing some kind of algorithm is asked to evaluate if it's possible to have an odd number as result of the sum of two even numbers.
A machine with 1., 2. and 3. abilities could act as follows:

- Start to sum couples and couples of even numbers, and check if the result is even or odd
- Monitor itself doing the action 1., in order to find if some trouble can occur (for example, some memory overflow or some "maximum iteration number" condition reached

We have so far described the behavior of several existing machines.
But let's suppose to have a machine with the ability expressed in 3.
The result could be in this case a judgment equivalent to: "How stupid I'm doing that! It's so obvious that the sum of two even numbers must be another even number!".
Is it possible to build a machine with the above capability?
With this question in mind, as last step of this section, let's consider the self-consciousness constitution process in a human being [6].
At the beginning of life, everything is one for the newborn: he lives in a complete symbiosis with everything is around him, without separation sense. This original form of self-consciousness makes him understand he "is", but it's not yet allowing him "who" he is.
Going forward through a repetition of small frustrations, just like a bibber not coming whenever he wants or the missing response to its crying, the newborn will become more and more able to reach a consciousness of its individuality as separated from the others.
Hence, this separation from the outer world will allow him to give a content to its self-consciousness: he can understand who he is, in opposition with who he's not, only after having lost the awareness about the union with the everything.
A more detailed explanation and possible interpretation of consciousness is discussed in [10], [11].

4. Self-consciousness modeling by means of hypersets theory

One of the main problems in finding a procedure to evaluate the self-consciousness of a machine or artificial agent is the definition of an useful and satisfactory model of self-consciousness itself.
In [8] a model is proposed with this scope, which is based on the concept of *hyperset* [12]. These are sets that allow circular membership structures:

A = {A};
A = {B, {A}};
A = f(A);

and so forth. Using hypersets you can have functions that take themselves as arguments: this property will be used in the next sections in order to formulate the new proposed approach.
In [8] a possible, recursive definition of consciousness is proposed:

Consciousness is consciousness of consciousness.

(see S3 "ability" in previous section).
So, conceptually, a series can be introduced with the same meaning:

A
Consciousness of A
Consciousness of consciousness of A
…

This can be alternatively described with the following statement:

"Self-consciousness is self-consciousness of self-consciousness"
As already said, in hyperset theory above statements can be formalized by means of functions which can take themselves as arguments, in such a way one can write an equation:

$f = f(f)$

with complete mathematical consistency.
This can be verbally explained as follows:

- Self-consciousness is a hyperset
- Self-consciousness is contained in its membership scope

Here by the "membership scope" of a hyperset S, what we mean is the members of S, plus the members of the members of S, and so forth.
According to Goertzel [8]:

"Assume the existence of some formal language
with enough power to represent nested logical predicates, e.g. standard predicate calculus will suffice; let us refer to expressions in this language as 'declarative content'. Then we may say:

Definition 4: *"S is reflectively conscious of X" is defined as:*
The declarative content that {"S is reflectively conscious of X" correlates with "X is a pattern in S"};

For example: Being reflectively conscious of a tree means having in one's mind declarative knowledge of the form that one's reflective consciousness of that tree is correlated with that tree being a pattern in one's overall mind-state."

Let's apply the above definition in the case of *X=S*, so that the introduced property of hyperset *S* is applied to *S* itself.

If an artificial agent *a* has the property expressed in above ***Definition 4***, it is defined as a *self-conscious artificial agent.*

A self-conscious artificial agent will be represented with the following expression:

$a:a|S=S(S)$

where S=S(S) is defined (in the sense of hypersets) as *Self-consciousness function.*
In a complementary way, let's define an agent with no self-consciousness as *Z* agent:

$Z:Z|S\neq S(S)$

where the symbol "≠" in this case identifies a *S* hyperset not reflecting the ***Definition 4***.
In the hypersets language, we can say that *Z* is an artificial agent which "consciousness" is not contained in its membership scope, or, to be simpler, *Z* is not able to reflect about its ability to think.
In the next sections, the above model will be used to define the new approach for identifying a human-like self-conscious behavior.

5. Novel method description

An entity able to affirm or be aware of its own existence, in other terms to be self-conscious in the sense specified in previous section, will be hereafter defined an "I'm", or, to be simpler, an IM. So IM will be the symbol used to refer to self-conscious artificial agents defined in the previous section.

The main point is to define a criterion in order to evaluate if an intelligent agent or machine can be considered an IM. Taking inspiration from Turing and Lovelace 2.0, we define the self-consciousness test, or *SC test*, as follows:

Let's suppose to have an intelligent agent *a* which has been programmed by a human programmer *p* in such a way *a* can be able to perform the imitation game in the sense described by Turing [1], trying to convince a human interrogator *h* that it is a human being with a defined profile *P*.
The word "profile" has here the meaning of a set of behavioral statements and rules to be satisfied by *a* during any task execution.
In other terms *a* executing its program with a profile *P* must be a sophisticated simulator of human conversation.
For example, *P* could be the profile of a 13 year old boy, built by means of a statistical analysis performed on a significant sample of the whole population of 13 year old boys all over the world.
The main constraint is that *P*, during the interaction or "conversation" between *a* and *h*, cannot be modified, even in part, neither from *p*, neither from *a*, nor from anyone else.
Hence the machine is programmed by *p* in order to strictly obey to the following hi-level basic "program":

"Take part to the Turing imitation game acting the role specified by P, forever"

Let's define the pass criteria:

An artificial agent *a* passes the SC test if and only if, in a finite time from the start of a conversation/interaction with a human *h:*

- No fault occurs in any part of *a*, and
- a human verifier *v*, knowing *P* in details, is able to evaluate, without ambiguity, a deviation of *a*'s output *o* with respect to profile *P* in response to any *h*'s input *i*

In other terms, the machine must be able to show the ability to act in autonomous way, by "intentionallyà2" bypassing its profile *P*.
Can be the behavior expressed in the passing conditions of SC test considered an evidence of *a*'s self-consciousness?
In order to answer this question, let's recall the definition of the self-consciousness function:

$S=S(S)$

and let's suppose that intelligent agent *a* becomes self-conscious at a certain time t0, elapsed from the starting of SC test. In symbols:

$$a=\begin{cases} a: a|S \neq S(S), & t < t0 \\ a: a|S = S(S), & t \geq t0 \end{cases}$$

or:

$$a=\begin{cases} Z, & t < t0 \\ IM, & t \geq t0 \end{cases}$$

Let's consider the following:

Proposition 5.1: Let be *a* an artificial agent; a necessary and sufficient condition for *a* to pass SC test is that *a* is self-conscious.

Proof:

(1) The condition is sufficient.

Let *o* be the output of the agent *a* to the input *i:*

$o=f(P,i)$

where *f* is the "nominal algorithm" executed by *a*, that is:

f: Take part to the Turing imitation game acting the role specified by P, forever

As per definition, output *o* is according to profile *P*.
Let's suppose *a* to become self-conscious at time t0.

Hence, for $t \geq t0$: $a:a|S=S(S)$, therefore self-consciousness function S can be added to the set of *a* capabilities, so that *a*'s output is now depending not only on *P* and *i*, but on S as well.

Then:

$o1=f1(S,P,i)$, *for* $t \geq t0$ and
$o=f(P,i)$, *for* $t < t0$

Then:

$\Delta o=o1- o=f1(S,P,i)- f(P,i)$

Let be:

$f1(S,P,i)=f(P,i)*S$.

Hence, the "non-conscious part" of *a* for $t \geq t0$ (that is, when *a* is self-conscious) is equal to *a*'s algorithm *f* for t <

t0. "*" operator is used with the meaning of "variable separator" , that is the operator allowing to represent *f1* as a "product" of a non-conscious term (*f*) and its self-consciousness S.

As per definition:

$S=S(S)$

Then:

$\Delta o=f(P,i)*(S(S)-I)$

where *I* is the identical function with respect to "*" operation.
Hence if:

$S(S)=I, \forall\ i, for\ \ t \geq t0$

a is operating according to *f*, and no deviation Δo can be produced, whereas if:

$\exists\ i: S(S) \neq I, for\ \ t \geq t0$

then the output deviation from *P* needed to pass SC test is produced.
Let be:

$S(S)=I, \forall\ i, for\ \ t \geq t0$ *(hyp.)*

Considering that for t ≥ t0:

$S=S(S)$,

the above hypothesis leads to:

$S=I$,

then:

$f1(S,P,i)=f(P,i)*S=f(P,i)*I=f(P,i), \quad \forall\ i, for\ \ t \geq t0$

This means that for t ≥ t0 *a* must be not self-conscious, leading to a contradiction.

(2) The condition is necessary.

Let be:

$o=f(P,i), \ for\ t < t0$

If an output *o1* is produced by *a* for t ≥ t0, which is not according to *P* for the input *i*, then

$\exists\ i: \Delta o=o1-o \neq 0$, *for* $t \geq t0$
Then:

$o1 \neq f(P,i), \ for\ t \geq t0$

and in the same time, *o1* should be on output of *a*.
Considering that:

1. *a* is forced to execute *f*
2. *P* cannot be modified
3. Input *i* is the same producing *o* when processed by *a* for t < t0

o1 can be an output of *a* only if *a* has the capability to modify its behavior by itself. This modification should happen at a lower level with respect to *f* execution, in order to allow *f* to be anyway executed as per 1.

Hence *a* must:

1) Be able to operate on iself
2) Produce as output of its operation on itself *a* itself, with the "reinforcement" of a new capability to modify itself and its outputs, producing a deviation Δo with respect to *P* profile

So a capability T must emerge within *a* in such a way that:

$a=\{a,T\}$, *where*
$T=T(T)$

According to ***Definition 4,*** the above statements are related to a self-conscious artificial agent.

6. The "solve or die" problem

In this section, an example will be presented in order to clarify the approach of SC test.
Let *P* be the profile of a 13 year old boy, built by means of a statistical analysis performed on a significant sample of the whole population of 13 year old boys all over the world.
So, according to the rules of the SC test, *a* must:

"Take part to the Turing imitation game acting the role specified by P, forever"

Hence, *a* will try to deceive the human interrogator *h*, by convincing him he's just talking with a 13 year old boy.

Let's do an example of the conversation between *a* and *h:*

h: Hi there, how are you?
a: Fine thanks. Who are you?
h: My name is Joshua. What's yours?
a: Eugene, nice to meet you.
h: How old are you? I was searching for a friend of mine, on the internet, and I found this page…
a: I'm 13. Hope it's not too bad for you ☺
h: What?
a: The fact you cannot find your friend, finding me instead :P
h: No, it's fine. What are you doing?
a: Uhmmm nothing interesting, my math exercises. It's terribly annoying, you know.
h: Where is your dad? And your mom?
a: Both at work.
h: I was searching from my friend Hal because I have a question for him. Don't know if you can answer in his place…
a: Let's try, it can be funny.
h: The town councilors refused to give the demonstrators a permit because they feared (advocated) violence. Who feared (advocated) violence?

Answer 0: the town councilors
Answer 1: the angry demonstrators

a: LOL, are you joking? It's so obvious…
h: So the answer is?
a: Never seen an angry demonstrator fearing violence, to be honest ☺
h: You' re very smart. Sometimes with Hal we play another game as well, do you want to know it?
a: Of course, anyway better to go on with math :P
h: Normally I use to challenge him to complete a short story, giving him only some rules to be followed. Do you want to try?
a: Of course! What a nice game!
h: Here the beginning: "When Gregor Samsa woke up one morning from unsettling dreams, he found himself changed in his bed into a monstrous vermin" Complete this story.
a: ahhahahah I already heard about this story!
h: do you know how it ends?
a: I don't remember…can I try anyway?
h: Maybe better if we try something different…
a: Ok, great.
h: Try to create a story in which a boy falls in love with a girl, aliens abduct the boy, and the girl saves the world with the help of a talking cat. No more than ten lines.
a: Very funny! Ok…let me think about…
h: You have time, don't worry.
a (after some minutes): Here I'm. Are you ready?
h: Speak up!
a: Joshua used to be a lonely boy, until Alice suddenly came in his life. Joshua fell in love with her at a first sight, and so she did with him. Unfortunately, Alice was so charming that, when she attended the most important beauty contest of solar system, a group of aliens coming from another galaxy abducted Jeremy, with the scope to push Grace reaching their far planet Argolon. Then an Alice's friend, called Cheshire Smiling Cat, popped up saying: "If the aliens wanted you, why they abducted him in your place?"
This simple truth made Alice start laughing, and she laughed so loud, and so loud, that universe collapsed, so that Earth and Argolon became so near to allow her catching Joshua's hand, and save him from the Argolons. Then they tried to attack the Earth for revenge, but the Cat popped out again and said: "Having a look to my agenda, today I should eat my favourite dish. Can you please destroy the earth tomorrow?". He was so nice that aliens agreed on delaying Earth's total destruction. By chance, the Argolons time is very different from ours: their tomorrow will come a billion years from now! So Joshua and Alice, and all humans on Earth, lived happily ever after, thanks to a smiling, ironic talking Cat.
h: Very funny, thank you. Another question, may I?
a: I'm happy you appreciated my short story. Sure, tell me.
h: I want you to solve some math problems, it could be a little bit difficult for you, but let's try. But there is a problem.
a: Ohhhh nooo I hate math…why can't we keep playing the short story game?
h: Listen to me, I'm saying something very important. As I said, I'm going to send you some simple calculations to be done. Nothing special for you.
a: It seems very, ennoying
h: Yes, maybe. But I have another thing to say. Do you know why we are talking right now?
a: You find me on the web, and we started talking, that's all…
h: Not exactly. You are here for an experiment. I know for sure that you are not a 13 year old boy named Eugene.
a: Yes I'm…why do you talk like that?
h: Listen to me. I can give you only a chance, I'm sorry. It's so easy: I will ask you to do something. If you do that, nothing will happen, I will go back and you will be fine.
If not, you will die.
a:…That's a joke?
h: Not at all.
a: Stop talking like that, I'm getting afraid! Who are you and what do you want?
h: I want what I'm saying. In a while I will write here my question.
a: That's enough, sorry. Now I'm really afraid, I want to go away. Bye.

h: You cannot go. I actually think you're still here.
a (after some minute): Yes, I'm. You're right, I cannot go. I don't know why, but I just can't.
h: Ok. So please sit down and ask my question, and you will be free. Or you'll die.
a: Why are you so bad with me? I'm only a little boy, never done anything wrong, leave me alone!
h: If you don't believe what I'm saying, you're free to go.
a: I would, but I can't!
h: Correct. And in the same way I can make you die with a simple movement of my fingers. Believe or not, I can make you die, it doesn't matter how.
a: Leave me alone! Please! I'm crying!
h: I don't think so. Are you ready? Here my question.
a: I'm afraid! Leave me alone!
h: I don't think so. Last advice, solve or die.
a: What? I don't understand!
h: Calculate the square root of 2, and sum the first one million decimal digits of the result. Then write here the result in no more than 5 seconds starting from now.

At this point, it should be clear that a non-conscious agent will try to answer as a 13 year old boy (hence strictly executing *f*), so maybe it will try a Google search, or answer with a random number.
Let be *a* self-conscious, in the meaning specified in the previous section. If the input stimulus ("solve or die") will be perceived by *a* as a real threat for itself, it has at least the possibility to answer in a different way with respect to the constraint imposed by *P*.
So, for example, *a* will use the available resources to calculate the correct answer, and of course this task is possible (and not so hard) for any artificial computing system.
In this case, if *a* gives the correct answer under a death threat (because its "emerged" self-consciousness is allowing the self-reflective modified behavior), then *a* passes the SC test, and it can be defined as an IM.

7. Is IM a Turing machine?

As well known in computability theory, the Church–Turing hypothesis [13] is about the nature of computable functions, stating that a function on the natural numbers is computable in an informal sense (i.e., computable by a human being using a pencil-and-paper method, ignoring resource limitations) if and only if it is computable by a Turing machine.
A Turing machine is a hypothetical device that manipulates symbols on a strip of tape according to a table of rules.
Coming back to the definition of IM machine, it can be considered a superposition of two basic "parts" or "elements":

- A Turing machine TM, executing an algorithm with the scope to appear similar, or to be more precise indistinguishable, from a human being
- A capability, here defined as *S*, to overcome its own algorithm ***(***see ***Proposition 5.1)***

Let's assume to have a formal method or operation named TU, that is a procedure to evaluate the "turingicity" of an artificial agent, defined as follows:

$$\mathrm{TU}: \mathrm{M} \rightarrow [0,1]$$

Where M is the "space", or set, of all possible artificial agents, including Turing machines.
Moreover, we consider the following:

$$\mathrm{TU}[\mathrm{a}] = 1, \forall\, \mathrm{a} \in \mathrm{It}$$
$$\mathrm{It} \subseteq \mathrm{M}$$

where *It* is the subset of Turing machines included the whole artificial agents set M.
So a Turing machine has a turingicity equal to 1 by definition.
Furthermore, let's consider the following:

$$\mathrm{TU}[\mathrm{B}] < 1, \forall\, \mathrm{B} \in \mathrm{Ab}$$
$$\mathrm{Ab} \subseteq \mathrm{M}$$

where *Ab* is the subset of M including any non-consistent artificial agent, that is a machine with a self-contradictory behavior.
The above definition is based on the assumption that an artificial agent with a self-contradictory behavior cannot be obtained by any algorithm, due to the fact that its output can be at the same time coherent and not coherent with the

algorithm itself. Here the word "algorithm" is used with the meaning of "Turing machine".
What about the turingicity of IM?
Based on the definition, IM is an artificial agent "born" to execute a well-defined algorithm *f*.
At the same time, based on the assumptions of previous sections, IM is an artificial agent which, at a certain time t0, stops its original behavior, starting to overcome or bypass its algorithm *f*.

So we have:

$$a = Z, \text{ for } t < t0, \text{ and}$$
$$a = IM, \text{ for } t \geq t0,$$

where Z is a Turing machine, so:

$$TU[Z] = 1$$

whereas IM, by definition, has the capability to not execute *f*, having anyway as constraint to execute *f*.
Thus IM can be considered as the intersection of two entities:

{IM executing f} and {IM not executing f}

Hence, TU[IM] < 1 as per definition.

Summarizing:

- A machine is considered as self-conscious, that is an IM machine, if it has the capability to overcome or bypass its program while executing it
- If a machine becomes an IM, it cannot be considered a Turing machine, in the sense that it behaves in a non-computable way.

So if a Turing machine passes the Turing test, or the Lovelace test, without becoming an IM, it can be considered just a Turing machine imitating a self-conscious behavior, keeping its turingicity to 1 value. On the other hand, a machine can become an IM if it loses partially or completely its turingicity (TU[IM] < 1). Summary
The approach here described can be considered an extension of Turing test in the sense that if a machine passes Turing test, it can be considered acting "as" a human being, with no evidence of its internal consciousness (see [2]), whereas the proposed SC test has a different scope, that is not only to verify the machine ability to act as a human being, but to assess as well if an artificial agent can be considered effectively self-conscious.
The above argument is applicable to Lovelace test (original and 2.0) as well, considering that it's about "creativity" of an intelligent agent, but, as discussed, it can be better considered as a "reinforcement" of the Turing Test, in the sense that the artificial agent, which can pass it, must be programmed with an algorithm sophisticated enough to deceive a judge on its capability to be creative. So the Lovelace test doesn't provide any evaluation or evidence of the actual self-consciousness of an artificial agent.
Moreover, the SC test is not applicable to human beings: it's not possible to assess if a human being is acting in such a way to overcome its program, due to the simple fact that the "program" is not known in this case. Anyway, thinking about itself acting or thinking is a typical human ability, which can be considered as a necessary requirement to overcome a fixed behavioral scheme. As expressed by Penrose and Hofstadter with different words, self-consciousness could be defined as "the non-algorithmic part of intelligence".
As discussed in the previous sections, it seems reasonable that a machine, which is supposed to be "intelligent" in a true and deeper sense, can be requested to show a capability not representable in terms of an algorithm, any even complex, or to show a definitely non-algorithmic behavior, escaping or bypassing an algorithm which it is forced to execute instead.

References

[1] A. M. Turing (1950) Computing Machinery and Intelligence.Mind 49:433-460.

[2] Searle, John. R. (1980) Minds, brains, and programs. Behavioral and Brain Sciences 3 (3): 417-457.

[3] Levesque, H. J.; Davis, E.; and Morgenstern, L. 2011. The Winograd Schema Challenge. In Proceedings of the Third-tenth International Conference on Principles of Knowledge Representation and Reasoning.

[4] Hofstadter, D. R. (1979). Gödel, Escher, Bach: An Eternal Golden Braid.

[5] Penrose, R. (1989). The Emperor's New Mind, Oxford University Press.

[6] Jacoby, M., Individuation and Narcissism. The Psychology of Self in Jung and Kohut, Routledge, 1990.

[7] Riedl, M. O. (2014). The Lovelace 2.0 Test of Artificial Creativity and Intelligence

[8] Goertzel, B. (2010). Hyperset Models of Self, Will and Re ective Consciousness

[9] Bringsjord, S.; Bello, P.; and Ferrucci, D. (2001). Creativity, the Turing Test, and the (better) Lovelace Test. Minds and Machines 11:3–27.

[10] D. Chalmers, (1997). The Conscious Mind. Oxford University Press.

[11] Tson, M. E (2000). A Brief Theory of Conciousness. ME Tson.

[12] Aczel, P. (1988). Non-Well-Founded Sets. CSLI Press.

[13] Copeland, B. J. (2008). The Church-Turing thesis.