

# In Circles:

## Rejecting the Lovelace Test

a criticism of "Creativity, the Turing Test, and the (Better) Lovelace Test"

<an anonymous philosophy student>

*Philosophy of Artificial Intelligence with Dr. S. Bringsjord  
Rensselaer Polytechnic Institute*

The thesis with which this paper is concerned is the Lovelace Test as presented by S. Bringsjord, P. Bello, and D. Ferrucci in their paper "Creativity, the Turing Test, and the (Better) Lovelace Test." First, I will explicate, in brief, why I believe that the Lovelace Test begs the question, in the philosophical sense of the phrase, of course. Next, I will clearly expound upon this view, logically elucidating my objection to the thesis. Then, I will give a possible objection to my position and then refute this objection. Finally, I shall close the paper with a summarising conclusion.

The Lovelace Test (*LT*) was designed and presented as a replacement for the original Turing Test (*TT*) (known otherwise as the "imitation game"), the more-sophisticated Total Turing Test (*TTT*), and the like. The primary objection to the *TT*, *TTT*, and variants (hence-forth referenced as the *TTF* – the Turing Test Family) was that they fail on account of Searle's Chinese Room Argument (*CRA*), allowing completely mind-less, symbol-manipulating machines to be considered creative and conscious, all based on "the strength of clever, but shallow, trickery" (Bringsjord, Bello, Ferrucci 2000). However, and although the *LT* does not fail where the *TTF* do, the *LT* is not a valid test for the reason that it begs the question – it presupposes weak artificial intelligence through "hidden" premises that are never made explicit in the paper, as will shortly be shown. On the basis of this logical fallacy, I will, ultimately, ask the reader to reject the *LT*.

The Lovelace Test is entirely invalid due to the fact that it presupposes weak artificial intelligence. The following are (paraphrased) Bringsjord's premises that are explicitly stated in the paper. Premise *E1* (explicitly-stated premise 1): Artificial agent *A*, designed by *H*, passes *LT* if and only if *A* outputs *o* and *A*'s outputting *o* is not the result of a fluke hardware error and *H* (or someone with *H*'s knowledge and resources) cannot explain *A*'s producing *o* by appeal to *A*'s architecture, knowledge-base, and core-functions. Premise *E2*: *A* is a standard Turing Machine (*TM*), a *TM* variant, or other equivalent digital-computing machine (i.e. not a chaotic analogue

neural network, etc.). Premise *E3*: An agent *Q* can be said to have originated something *R* if and only if *Q* single-handedly created *R*, knowing full-well what it was doing and intending consciously to do so. About these, I suspect, there will be no argument. However, what may be debated, along with the truth of premise *E3*, are the premises used but stated either tacitly or not at all. These are as follows. Premise *I1* (implicitly-stated premise 1): The human mind cannot be the subject of the *LT*. Premise *I2*: the human mind is (or, at the very least, can be) creative. The premise *I1* is derived from the facts (along with knowledge that the authors of the paper, particularly Bringsjord, believe that humans cannot possibly be computers) *a*) that the creator of humans, *H<sub>humans</sub>* is unknown or, at best, unreachable for participation in the test *b*) that it is not, even in principle, possible to determine whether or not “fluke hardware errors” occur during human cognition *c*) that the inner workings of the human brain/mind are not fully known. For the reasons enumerated (viz. *a*, *b*, and *c*), humans are invalid candidates for the test. This, I suppose, could be mitigated by the creator of humans (either evolution or God or some other force) serving as the creator, *H*, but no such being has come forward for inclusion in this test. Premise *I2* is derived from common-sense.

In further explication, the *LT* fails to be valid for the fact that this “escape-route” for humans being able to pass the test indicates that the *LT* has no basis as a criterion for being granted the appellation “creative.” Restated, since the test cannot be applied to a being that, by premise *I2* (a premise to which I agree fully), is known to be creative, it is absurd. It is evident, of course, that the only reason for which the *LT* is inapplicable to humans is the presumed “fact” that human cognition is uncomputable. Should it be possible, through some means or other, to determine a complete algorithm for a particular human *H\**'s cognitive processes, the *LT* could be applied to *H\** and, thus, *H\**, by definition of the test, would be found to be uncreative since 1) *H\** would produce output  $\sigma^*$ , 2) *H\** doing so would

not be the result of hardware failures and 3) anyone knowing  $H^i$ 's algorithm would be able (by definition of an algorithm) to trace through the steps necessary to produce  $\sigma^*$ , given  $H^i$ 's initial state and inputs. Hence, it would seem, the entire strength of the  $LT$  argument lies in the fact that an algorithm for human thought has not yet been discovered and, by all current technologies, it is not humanly possible that any shall be scientifically-derived, stumbled upon, or otherwise discovered in any nearly-distant future. The argument, although based on fallacy, will appear to stand as truth until this falsity of logic in  $LT$  is discovered, by those giving it credence, as it has been discovered in the process of writing this paper.

It seems that premise  $E3$  is also subject to attack. Many of the innovations which have been attributed to human "creativity" or "ingenuity" are actually the result of accidents. Furthermore, I would argue that this is of no consequence, the innovations being still the product of the human mind. As examples, consider "silly putty," the "slinky" toy, and (on the more useful side) Vulcanised rubber. These were all the product of "happy accidents" on the part of the inventors (as I recall, silly putty was the result of a glue-developer's accident, slinky toys the result of a spring-designer's error, and Vulcanised rubber the outcome of Goodyear erroneously dropping a piece of rubber into a stove), but they are attributed and praised as are any other inventions. Making the most of an accident is one of the hallmark signs of creativity – taking and experiment-gone-wrong or an accident and turning the resultant product into something useful is one of the most difficult and creativity-requisite functions of an innovator. Likewise, many a great painting has been crowned by a misplaced brush-stroke and many a building made famous by a design flaw (i.e. the Leaning Tower at Pisa). Hence, I roundly reject conditionally  $E3$  as a qualification for creativity.

Of course, I anticipate Bringsjord (and, most likely, others) raising objections to my objection. One such objection that may be considered is that presented in the last paragraph of the paper which is the topic of

scrutiny, namely that of autonomy, that the agent in question of the test must, in some very genuine sense, be autonomous in its operation. I respond to this by stating that this autonomy, possibly agent-causation, is impossible, given the constraints of the test: the test may be applied only to *TM*s, which, by the very definition of a Turing Machine, are deterministic (or, if non-deterministic, are no more powerful than an equivalent deterministic *TM*, as has been shown in computational theory) and completely without "freedom" in any sense. Hence, it is by definition that nothing which is the subject of the test will ever pass the test. Thus, also, it is that the test is entirely worthless, for what good is a test whose outcome is known always to be false? No *TM*-based system can ever be creative by definition of *LT*.

In conclusion, I do not have a good test for creativity, personhood, the possession of an inner mental life, or cognition. I, ultimately, agree with Bringsjord that computers cannot be creative in the way that are humans, but am unconvinced by the *LT* argument. I do not think that such a test can, in fact, be mathematised or algorithmatised, for these are concepts far too abstract for such formalisms. Rather, and much to my dismay as a philosopher-mind, we must (defeasibly) accept the personhood of a being based solely on behaviourism, an acceptance which may be later rescinded based on the discovery of a being's computational nature. Or, as it seems more useful to me, it does not matter what the nature of one's mind; for me it suffices that someone pass the "imitation game" (*TT*), for how do I really know that you have an inner mental life? Only based on the outputs that you provide me: conversation, gestures, expressions, etc. If Pollock is successful in his pursuance of the Oscar Project's ultimate goals, then, indeed, any dissension to his approach will be, as he says, "passé" – at least, that is, until a better test of personhood or creativity is developed.