The Lovelace 2.0 Test of Artificial Creativity and Intelligence

Mark O. Riedl

School of Interactive Computing; Georgia Institute of Technology

riedl@cc.gatech.edu

Abstract

Observing that the creation of certain types of artistic artifacts necessitate intelligence, we present the Lovelace 2.0 Test of creativity as an alternative to the Turing Test as a means of determining whether an agent is intelligent. The Lovelace 2.0 Test builds off prior tests of creativity and additionally provides a means of directly comparing the relative intelligence of different agents.

Introduction

Alan Turing proposed the Imitation Game—later referred to as the Turing Test—as a lens through which to examine the question of whether a machine can be considered to think (Turing 1950). The Turing Test was ever meant to be conducted; indeed many practical methodological details are left absent by Alan Turing. Regardless of Turing's intent, the Turing Test has been adopted as a rigorous test of the intelligence capability of computational systems. Occasionally, researchers make claims that the the test has been passed. The most recent claim involved a chatbot using simple template matching rules. One of the weaknesses of the Turing Test as a diagnostic tool for intelligence is its reliance on deception (Levesque, Davis, and Morgenstern 2012); agents that are successful at the Turing Test and the closely related Loebner Prize Competition are those that fool human judges for short amounts of time partially by evading the judges' questions.

A number of alternative tests of intelligence have been proposed including Winograd Schemas (Levesque, Davis, and Morgenstern 2012), question-answering in the context of a television show, and a robot that gives a talk at the TED conference. Bringsjord, Bello, and Ferrucci (2001) proposed the Lovelace Test, in which an intelligent system must originate a creative concept or work of art. For certain types of creative acts, such as fabricating novel, fictional stories, it can be argued that a creative computational system must possess many of the cognitive capabilities of humans.

In this paper, we propose an updated Lovelace Test as an alternative to the Turing Test. The original Lovelace Test, described in the next section, is thought to be unbeatable.

The new Lovelace Test proposed in this paper 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.

Background

Hartree (1949) quotes Ada Lovelace: "the Analytical Engine has no pretensions to *originate* anything. It can do *whatever we know how to order it* to perform" Turing (1950) refutes the charge that computing machines cannot originate concepts and reframes the question as whether a machine can never "take us by surprise."

The original Lovelace Test (Bringsjord, Bello, and Ferrucci 2001) 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.

Computational creativity is the art, science, philosophy, and engineering of computational systems that, by taking on particular responsibilities, exhibit behaviors that unbiased observers would deem to be creative. There are no conclusive tests of whether a computational system exhibits creativity. Boden (2004) proposes that creative systems be able to produce artifacts that are *valuable*, *novel*, and *surprising*. Unfortunately, it is not clear how to measure these attributes. Boden describes surprise, in particular, as the experience of

realizing something one believed to be highly improbably has in fact occurred. Automated story generation is the fabrication of fictional stories by an artificial agent and has been an active area in computational creativity. The strong story hypothesis (Winston 2011) states that story understanding and story telling play a central role in human intelligence.

The Lovelace 2.0 Test

We propose a test 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.

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. For example: "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 ability to correctly respond to the given set of constraints C is a strong indicator of intelligence.

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. Aesthetic valuations are not considered. We suggest that the judge be allowed to repeat the test any number of times with different t and C.

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.

With a little bit of additional methodology, the Lovelace 2.0 Test can be used to quantify the creativity of an artificial agent, allowing for the comparison of different systems. Suppose there is a set H of human evaluators, each of which performs a sequence of Lovelace Tests, $k = 1...n_i$, such that $|C_k| = k$ and n_i is the first test at which the agent fails to meet the criteria given by evaluator $h_i \in H$. That is, each evaluator runs the Lovelace Test where the kth test has k constraints and stops administering tests after the first time the agent fails the test. The creativity of the artificial agent can be expressed as the mean number of tests passed: $\sum_i (n_i)/|H|$. With a sufficiently large $|H|$, one should get a good idea of the capabilities of the system.

The Lovelace 2.0 Test is a means of evaluating the creativity of an entity with respect to well-defined types of artifacts. The proposed test can also act as a test of intelligence in the case of types of artifacts that require human-level intelligence. Consider a limited form of the test: the generation of fictional stories. Fictional story generation requires a number of human-level cognitive capabilities including commonsense knowledge, planning, theory of mind, affective reasoning, discourse planning, and natural language processing. A story generator is also likely to benefit from familiarity with, and able to comprehend, existing literature and cultural artifacts. Currently, no existing story generation system can pass the Lovelace 2.0 Test because most story generation systems require *a priori* domain descriptions. *Open story generation* partially addresses this by learning domain knowledge in a just-in-time fashion (Li et al. 2013), but cannot yet comprehend and address complex constraints.

The Lovelace 2.0 Test is designed to encourage skepticism in the human evaluators. Regardless of whether the human judge is an expert in artificial intelligence or not, the evaluator is given the chance to craft a set of constraints that he or she would expect the agent to be unable to meet. Thus if the judge is acting with the intent to disprove the intelligence, the judge should experience an element of surprise if the agent passes the test. The ability to repeat the test with more or harder constraints enables the judge to test the limits of the agent's intelligence. These features are at the expense of a halting function—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. In the spirit of the Imitation Game, the Lovelace 2.0 Test asks that artificial agents comprehend instruction and create at the amateur levels.

References

[Boden 2004] Boden, M. 2004. *The Creative Mind: Myths and Mechanisms.* Routledge, 2nd edition edition.

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

[Hartree 1949] Hartree, D. 1949. *Calculating Instruments and Machines.* University of Illinois Press.

[Levesque, Davis, and Morgenstern 2012] Levesque, H.; Davis, E.; and Morgenstern, L. 2012. The Winograd schema challenge. In *Proc. of the 13th International Conference on the Principles of Knowledge Representation and Reasoning.*

[Li et al. 2013] Li, B.; Lee-Urban, S.; Johnston, G.; and Riedl, M. O. 2013. Story generation with crowdsourced plot graphs. In *Proc. of the 27th AAAI Conference on Artificial Intelligence.*

[Turing 1950] Turing, A. 1950. Computing machinery and intelligence. *Mind* 49:433–460.

[Winston 2011] Winston, P. H. 2011. The strong story hypothesis and the directed perception hypothesis. In *Proc. of the 2011 AAAI Symposium on Advances in Cognitive Systems*.