Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 123 (2018) 512–521

Procedia Computer Science

www.elsevier.com/locate/procedia

8th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2017

A Conceptually Different Approach to the Empirical Test of Alan Turing

Ilya Vorobiev[1] and Alexei V. Samsonovich[1,2]

[1]*National Research Nuclear University "MEPhI" (Moscow Engineering Physical Institute)*
Kashirskoe Shosse 31, Moscow, 115409, Russian Federation
[2]*Krasnow Institute for Advanced Study, George Mason University, Fairfax, VA 22030, USA*
kotolin3000@yandex.ru, asamsono@gmu.edu

Abstract

The test for general human-level intelligence proposed by Alan Turing at the dawn of computers remains to be the best-known ultimate challenge for Artificial Intelligence today, if understood more generally than the originally proposed paradigm, despite all its controversies and weaknesses. Yet, a better-posed challenge is needed in the era of new technological capabilities and ambitions: a challenge that can drive research toward practically vital achievements. In this work, a new general challenge based on a Turing-test-like paradigm is formulated, together with a roadmap based on it. It is argued that the proposed roadmap is vital for the integration of future robots and virtual agents into the human society.

Peer-review under responsibility of the scientific committee of the 8th Annual International Conference on Biologically Inspired Cognitive Architectures

Keywords: Turing test, human-level AI, emotional intelligence, believable agents, metrics & challenges, roadmap

1 Introduction

In the 20th century, the English mathematician Alan Turing actively dealt with the question of the possibility of Artificial Intelligence (AI) and what features it should possess, as described in his work "Computing machinery and intelligence" [1]. In fact, he was the first to address this issue and, accordingly, set out a certain basis that the scientific society operated on in the future. The Turing test is one of the most famous and controversial experimental paradigms known to this day even to people not close to science. Despite its harsh criticism [2,3], it still remains associated with our understanding of the ultimate goal of AI research. But, is it really the best we can do in terms of setting the ultimate goal – or, is there an alternative? Here we entertain one potentially promising view.

The standard interpretation of the Turing test is as follows. A human evaluator interacts with two other participants, one of whom is a human and another is an artifact, such as a computer program. Based on the interaction, the evaluator must determine who she is talking to: a person or an artifact.

Peer-review under responsibility of the scientific committee of the 8th Annual International Conference on Biologically Inspired Cognitive Architectures
10.1016/j.procs.2018.01.078

The task for either of them is to convince the evaluator that they are a human person, and the other participant is an artifact. If the judgment is made at a chance level, then the artifact is said to pass the Turing test and in this sense can be called a human-like artificial agency. While the originally proposed interaction paradigm included only text questioning by the evaluator with certain restrictions, today more rich versions are considered, including free behavior in full-fleshed robotic embodiments [3]. The most general paradigm described above can be summarized by a diagram (Figure 1).

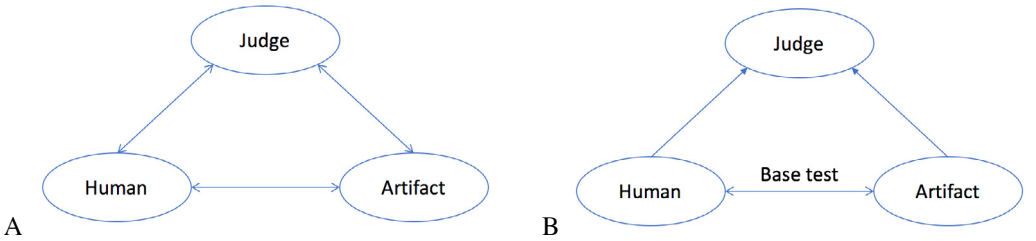

Figure 1: Generalized paradigms of the Turing test. Here "Judge" is the human evaluator, while in principle this role could be played by an artifact. A: Bi-directional interactions occur among all three participants. B: Bi-directional interactions occur only at the base level (Human-Artifact), while Judge gives no information to other participants during the test.

We eschew any rigorous analysis of many problems associated with this test (see, however, [2,6]). Instead, we point to only one of them and to a possible way of its mitigation. A successful session of a Turing test cannot guarantee a success in another, follow-up session, and could result from a random outcome. Therefore, a successful Turing test does not establish a new general fact. Indeed, the success in this case is a negative result (telling us that the null hypothesis, H0, cannot be rejected), which possibly can be overturned by another test based on a sufficiently large sample.

Imagine, on the other hand, that somebody builds an artifact that will be judged to be a human more frequently than an average human, based on a certain limited version of the Turing test. Then, once this fact is reliably established empirically (H0 is rejected), it will remain a fact.

2 Preliminary Analysis

At first, the idea that an artifact can be in any sense "more human-like" than any of the humans strikes as a logical contradiction. Indeed, how anything can be more similar to something than the thing itself? This impression, however, results from a misunderstanding of the idea. Recall that in the Turing test one does not measure the overall similarity of an artifact to a human in terms of a multitude of objective metrics and characteristics. Instead, the test relies on one measure: participant's subjective judgment of the believability, or human-likeness of an artifact, expressed as a "yes-no" answer to the question of whether a certain entity is a human or not. The goal in the Turing Test challenge is to make probabilities of the positive answer statistically equal, for a human and for a computer. In contrast, the goal in the Overman Challenge (this is how we propose to name it) is to make the probability of the positive answer significantly higher for a computer, as compared to a human. Is it logically possible? Yes, if an artifact can exhibit behavior that is typical of a human who scores higher than average on believability. To make this idea precise, it is necessary to define certain terms and conditions first.